Infectious disease modelling and causal inference

Jon Michael Gran

Dissertation for the degree of PhD

Department of Biostatistics
Institute of Basic Medical Sciences
Faculty of Medicine
University of Oslo
Norway

Oslo, September 2010

Series of dissertations submitted to the
Faculty of Medicine, University of Oslo
No. 1087

ISBN 978-82-8264-061-9

Cover: Inger Sandved Anfinsen.
Printed in Norway: AIT Oslo AS.

Produced in co-operation with Unipub.
The thesis is produced by Unipub merely in connection with the
thesis defence. Kindly direct all inquiries regarding the thesis to the copyright
holder or the unit which grants the doctorate.

Contents

Acknowledgments

First and foremost, I would like to thank my supervisor Odd O. Aalen for his excellent guidance and support.

I would also like to thank my other co-authors for all their efforts and advice along the way. In particular, I would like to thank my colleague Kjetil Røysland for the collaboration over the past three years.

I would like to thank the Research Council of Norway (grant number 17062/V30) and the Centre for Biostatistical Modelling in the Medical Sciences (BMMS) for funding.

Many thanks to my colleagues at Department of Biostatistics for creating such a great academic and social environment.

Finally, a special thanks to my family and friends, and especially Veronica, for all the encouragement and for making these four years so enjoyable.

Oslo, September 2010
Jon Michael Gran

List of papers

Paper 1

Gran JM, Wasmuth L, Amundsen EJ, Lindqvist BH, Aalen OO. Growth rates in epidemic models: Application to a model for HIV/AIDS progression. *Statistics in Medicine* 2008; **27**(23):4817–4834. DOI: 10.1002/sim.3219

Paper 2

Gran JM, Iversen B, Hungnes O, Aalen OO. Estimating influenza-related excess mortality and reproduction numbers for seasonal influenza in Norway, 1975–2004. *Epidemiology and Infection* 2010; **138**(11):1559–1568. DOI: 10.1017/S0950268810000671

Paper 3

Gran JM, Røysland K, Wolbers M, Didelez V, Sterne J, Ledergerber B, Furrer H, von Wyl V, Aalen OO. A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine* 2010; 1–12. [E-pub ahead of print]. DOI: 10.1002/sim.4048

Paper 4

Røysland K, Gran JM, Ledergerber B, von Wyl V, Young J, Martinussen T, Aalen OO. Analysing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study. 2010; 1–22. [Manuscript].

1 Introduction

Over the past two decades there has been a new and increased interest in the concept of causality in statistics. This interest has lead to various new approaches for defining and studying causality in statistical terms. In medical sciences, studies are typically driven by questions that are causal in nature, for example questions of treatment effects or the effect of some other intervention. Causality can, however, also be about understanding mechanisms, addressing questions about possible mediating variables and their effects. For example, considering which variables the treatment effect is mediated through. Answering any such questions about causality systematically requires extensions to standard statistical methods.

Infectious disease modelling is a field that has also received an increasing amount of attention over the last decades with the spread of HIV, the recent swine flu pandemic, the threat of avian flue and many others. Infectious diseases have re-emerged as a threat to western countries; and there has also been an increasing focus on Third World medical problems. As medical research in Western countries traditionally has had chronic diseases such as heart disease and cancer as the major areas of focus, new statistical models are required to understand the spread of infectious diseases.

The main aims of this thesis are to study approaches for estimating the causal effect of treatment on survival from HIV/AIDS using data from observational studies, and the concepts of epidemic growth rates and dynamic modelling of infectious diseases. These aims are met by studying methods for estimating the causal effect of treatment, and addressing the concept of direct and indirect effects by dynamic path analysis, using data from the Swiss HIV Cohort Study. Epidemic growth rates and reproduction numbers are studied together with a dynamic model for HIV/AIDS progression and models for estimating influenza related excess mortality using Norwegian surveillance data.

A recurrent theme in this thesis is infectious disease, and the methods

discussed are all applied in either a HIV/AIDS or influenza setting. However, the methods cover different areas of statistical and mathematical modelling; the two most important are infectious disease modelling and causality. There are four papers included in this thesis, and even though **Paper 3** and **Paper 4** address causal inference directly, **Paper 1** and **Paper 2**, considering infectious disease modelling, complement these last two papers with regard to causality in infectious diseases. **Paper 1** and **Paper 2** illustrate how infectious disease modelling is strongly connected to a causal way of thinking. There are obvious similarities between mechanistic causality as discussed in **Paper 4**, and modelling infectious diseases using dynamic multi-stage models, as in **Paper 1**. **Paper 2** is an example of how questions in medical science are causal in nature, and how these questions can be addressed using public health surveillance data.

In **Paper 1** we study the use of epidemic growth rates and reproduction numbers, with application to a multi-stage model for HIV/AIDS progression. In **Paper 2** we estimate influenza related excess mortality, epidemic growth rates and reproduction numbers using Norwegian surveillance data. In **Paper 3** we introduce a method for estimating the causal effect of treatment using data from a HIV/AIDS cohort study, while in **Paper 4** we build on some of the ideas from the previous paper to analyze the same type of data using dynamic path analysis. The four papers follow at the end of this thesis.

The outline for the rest of the thesis is as follows: Section 2 gives an introduction to infectious disease, focusing on HIV/AIDS and influenza in particular. Modelling of infectious disease, HIV/AIDS progression, influenza related excess mortality, reproduction numbers and growth rates are discussed in Section 3. Causality and causal modelling are discussed in Section 4, in particular with respect to counterfactual causality and to methods for making causal inference from observational studies. Direct and indirect effects are discussed in Section 5, emphasizing graphical models and dynamic path analysis. The four papers are summarized in Section 6, while a final discus-

sion is found in Section 7.

2 Infectious diseases

Infectious diseases are diseases caused by a transmissible agent replicating in an infected host. New infections occur when susceptible hosts are exposed to, and then acquire the agent. Agents can transmit from one host to another, creating a chain of transmissions in the population; this is the most distinctive feature of infectious disease epidemiology [1].

There are many reasons why infectious disease has received more attention in recent years: influenza and HIV have already been mentioned, while other examples are the occurrence of resistant tuberculosis, the SARS epidemic, the fear of bio-terrorism, multi-resistant microbes in hospitals and various sexually transmitted diseases [2]. However, as the methods in this thesis are all applied to either data on HIV/AIDS or influenza, the following discussion will be limited to these two infectious diseases.

2.1 HIV/AIDS

Human immunodeficiency virus (HIV) is a virus that attacks the human immune system by infecting cells such as the helper/inducer T cells. These cells are also designated as CD4 cells. The stage of HIV infection is typically measured by the CD4 cell count and by the viral load (the number of copies of HIV RNA in the blood). The most severely affected patients are prone to repeated opportunistic infections and neoplasms, characterized as the *acquired immune deficiency syndrome* (AIDS) [3]. Since the discovery of HIV and AIDS in the early eighties, the disease has spread worldwide at an alarming rate. The pandemic was in 2002 estimated to have infected 42 million people, increasing at a rate of 5 million new infections per year, and causing 3 million deaths per year – the majority were in the sub-Saharan Africa [4]. In most western countries, the HIV epidemic is mostly associated with certain risk

groups, such as men who have sex with men and injecting drug users. Recent treatment, using *highly active antiretroviral treatment* (HAART), and often referred to as just *antiretroviral treatment* (ART), has shown a substantial reduction in disease progression among infected patients [5], but a vaccine has yet to be discovered.

Historically, HIV/AIDS surveillance in the western countries focused on monitoring the AIDS incidence, and even after the HIV test became available, reports of HIV incidence were slow to be adopted. However, after the introduction of HAART, disease progression became less predictable and AIDS incidence no longer represented transmission trends. HIV case-reporting therefore became more urgent. HIV prevalence was typically monitored through surveys among different risk groups [6]. Today there are many huge multicenter cohort studies, national and international collaborations, gathering relevant information about HIV infected individuals. Examples include the Multicenter AIDS Cohort Study (MACS) [7] in the USA and the Swiss HIV Cohort Study [8].

The methods in **Paper 1**, **Paper 3** and **Paper 4** in this thesis are all applied to HIV/AIDS.

2.2 Influenza

Influenza is an infection of the respiratory tract caused by influenza viruses. The viruses are transmitted by contact, droplet and airborne transmission, and are highly transmissible in high density communities [9]. The infection is typically characterized by sudden-onset fever, dry cough, nasal congestion, headache, muscle pain, weakness, and loss of appetite [10]. Influenza is usually self-limited with recovery in 2–7 days. Pneumonia is a common complication. Most patients make a full recovery, but the severity of influenza epidemics vary with virus type, and for certain high risk groups influenza can be a deadly disease. The highest mortality is found among the elderly and patients with underlying severe diseases, and for these risk groups annual in-

fluenza vaccines are recommended in many countries [11]. Seasonal influenza usually causes annual outbreaks of varying length and severity during the winter months in the northern hemisphere. Occasionally, new virus variants appear, to which few or no-one is immune, possibly causing an influenza pandemic. There were three big influenza pandemics in the last century: the Spanish flu in 1918–19, the Asian flu in 1957–58 and the Hong Kong flu in 1968–70. More recently, we had the so-called swine flu pandemic emerging from Mexico in 2009.

Most developed countries monitor influenza activity by measuring the number of patients seeking health care, the virologically confirmed cases or by other markers for *influenza-like illness* (ILI). In Norway from 1998, weekly numbers on influenza have been collected from sentinel practices using the number of ILI cases per total number of consultations. Mortality due to influenza is, on the other hand, harder to discern, and influenza is rarely recorded as the cause of death in Norway [11].

The methods in **Paper 2** in this thesis are applied to Norwegian ILI data.

3 Infectious disease modelling

The notion of *infectious disease modelling* is most often used for infectious disease transmission models, compartment models or dynamic systems explaining and predicting the movement of infection through a population over time [1]. Even though there is a long history of mathematical modelling in epidemiology, going back to Bernoulli in the 18th century, the dynamic system approach first became popular in the 1920s. Since then, the methodology surrounding this type of models has been exposed to numerous conceptual and technical developments [4]. The models are used to predict the course of epidemics and to investigate the effect of change in model parameters. Changes in model parameters can represent change of risk-seeking behavior, the introduction of vaccines, medication or any other countermeasure.

Infectious disease modelling has become a huge field, and the methodology covered in this thesis will be limited to certain groups of models. We will consider the type of compartment models mentioned in the introduction of this section, which are used to model HIV/AIDS progression in **Paper 1**. Other types of models are models for estimating excess mortality using epidemic surveillance data and data on overall mortality, as studied in **Paper 2**. These latter models are somewhat on the side of what is usually denoted infectious disease modelling, as they do not model the spread or progression of disease as the compartment models above. Instead, they are based on fitting models to observed epidemic curves using regression techniques. Other important concepts within infectious disease modelling are reproduction numbers and epidemic growth rates. Both **Paper 1** and **Paper 2** discuss the estimation and use of these quantities.

Examples of other areas in infectious disease modelling, not covered in this thesis, are expansions of the mentioned compartment models into multi-host models, models with stochastic dynamics, spatial models and network models. For an introduction to these topics, see for example Keeling and Rohani [4].

3.1 Compartment models

Generally, transmission models for infectious disease consider members of the population to fall into *compartments*, *stages* or *states* such as 'susceptibles', 'infectious' and 'recovered' [1]. This example of compartments is used in the famous *SIR model* by Kermack and McKendrick from the 1920s [4]. Simple transmission models for infectious disease use only two or three compartments, but many more compartments can be used in more complex situations. An example of a basic SIR model is illustrated in Figure 1. Here, S, I and R denote the susceptible, infectious and recovered part of the population. Correspondingly, the notation $S(t)$, $I(t)$ and $R(t)$ is used for the number of individuals in each of the three compartments. $\lambda(t)$ and γ are

transition rates, specifically the rate of infection and the rate of recovery. The infection rate $\lambda(t)$ is dependent of the number of susceptible individuals in the population, usually by $\lambda(t) = k \cdot s(t)$, where k is the number of new infections transmitted by one individual per unit time and $s(t)$ is the proportion of susceptible individuals in the population at time t.

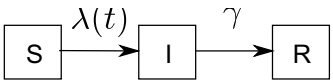

Figure 1: An example of a simple SIR model with transition rates $\lambda(t)$ and γ.

The SIR model shown in Figure 1 can be modeled using the ordinary differential equations

$$\begin{aligned} \frac{dS}{dt} &= -\lambda(t)I(t), \\ \frac{dI}{dt} &= \lambda(t)I(t) - \gamma I(t), \\ \frac{dR}{dt} &= \gamma I(t). \end{aligned}$$

Despite their simplicity, these differential equations cannot be solved explicitly and are therefore solved numerically [4]. The models can of course be extended and altered in various different ways, for example by including *demography* (individuals entering and leaving the system with certain rates), other dependencies between compartments or by including more compartments. For further reading on the SIR model and related models, see **Paper 1** and also Keeling and Rohani [4] or Diekmann and Heesterbeek [12].

3.1.1 A model for HIV/AIDS progression

One extension of the simple compartment models discussed in the previous subsection is described in **Paper 1**, as a model for HIV disease progression among homosexual men in England and Wales. See Figure 2 for an

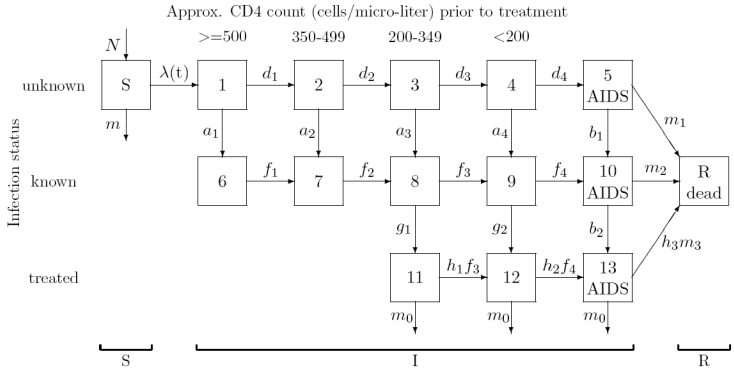

Figure 2: A generalized SIR model for HIV/AIDS progression.

illustration (figure reprinted from **Paper 1**).

The model in Figure 2 is a generalized SIR model [13] and, compared with the model in Figure 1, it has been extended with demography and the use of a multi-stage Markov model to replace the single infectious compartment. The added demographic parameters include N, m and m_0. N is the number of individuals entering the susceptible stage per unit time, while m is rate to which the susceptibles die, or leave the system uninfected, and m_0 is the background mortality rate. Note that a background mortality rate, such as m_0, should generally affect all states where individuals are still alive. However, for simplicity, m_0 is only added to state 11, 12 and 13 in the model in Figure 2. The rate m_0 is small compared to the other transition rates and it makes no practical difference whether to include them. Including such rates could, on the other hand, be essential for other exposure groups, such as injecting drug users [14].

The other parameters $\lambda(t)$, a_i, b_i, d_i, f_i, g_i and m_i for all i are transition rates, and h_1, h_2 and h_3 are factors decelerating, or modifying, certain transition rates. The model is based on the Markov model in Aalen *et al.* [18],

and the parameters are based on previous estimated rates and on published studies of HIV/AIDS. See **Paper 1** for further details.

3.2 Estimating influenza-related excess mortality

Another type of model that can be applied to infectious diseases are models for estimating the *excess mortality* attributable to disease using surveillance data on disease incidence and numbers of overall deaths. In **Paper 2**, such models are used to estimate the excess mortality due to influenza. As in this paper, the discussion in the following section will focus mainly on applications to influenza. These models could however also be used to estimate other types of excess mortality. For example, they are commonly used to estimate excess mortality related to heat waves as in Fouillet *et al.* [15]. Generally, the models will not only apply to excess mortality, but also to other equivalent situations, such as estimating excess hospitalization [16].

Excess mortality is usually defined as the difference between the observed number of deaths and the expected number of deaths, where the latter is the expected number of deaths without the occurrence of influenza. There are various ways of quantifying this number of expected deaths, but they can generally be divided into three types of models.

In the most simple *classic epidemic-threshold models*, the excess mortality is defined as all deaths exceeding a seasonal baseline threshold, found for example by averaging over periods without influenza outbreaks.

Models based on *cyclic regression* or *time series* calculate this threshold in a more advanced manner, modelling seasonal variation using parametric models. Typically, the parts of the mortality and influenza time series where epidemics are meant to have occurred are excluded from the estimating process [17].

Other models are based on *Poisson regression*, *negative binomial regression* or other related regression models. Unlike the cyclic regression and time series models, they model the mortality using influenza activity as an

explanatory variable, together with other important explanatory variables. The models used in **Paper 2** fall into this category. Using these models the excess mortality is estimated as the difference between the observed and the predicted mortality from the regression model, leaving out the influenza contribution.

Beside the aforementioned differences between these three types of models, they also vary significantly within these categories. The differences will typically lie in the model specification, the type of marker for influenza activity and other variables affecting mortality, which additional covariates are being included, reasoning surrounding what periods should be excluded from the estimating process, whether weekly data are used or data with another resolution and so on. All of these aspects could affect the excess mortality estimate. The principal advantage of using models based on Poisson or negative binomial regression is that other competing causes for excess mortality, such as the incidence of other diseases or factors such as temperature, can be included as covariates. Also, modelling the seasonality using markers means that parametric shapes of seasonality do not need to be assumed. However, if data only are available for short time periods, the ability to add such parametric assumptions would be an advantage.

For further discussion and references, see **Paper 2** as well as Nunes *et al.* [17].

3.3 Reproduction numbers and epidemic growth rates

A common goal when modelling infectious disease is to quantify the growth of an epidemic, usually by estimating reproduction numbers. These are important quantities because they give a better understanding of the magnitude and the potential course of an epidemic. For example, calculating reproduction numbers from compartment models can give information about what type of countermeasures would be sufficient for preventing outbreaks. Reproduction numbers and epidemic growth rates are among the topics in **Paper 1**

and **Paper 2**.

Generally, a *reproduction number* is the average number of secondary infections caused by one infected individual during the infectious period.

The term *basic reproduction number* (R_0) is used for the reproduction number of one individual introduced into a completely susceptible population. In a homogeneous population with random mixing, where c is the expected number of contacts per unit time, β is the probability that a contact between a susceptible and an infected leads to transmission and D is the average length of the infectious period, R_0 can be calculated as [12, 13]

$$R_0 = c\beta D$$

Alternatively, for the basic SIR model in Figure 1, using the corresponding notation, R_0 is given by

$$R_0 = \frac{k}{\gamma}. \tag{3.1}$$

For basic SIR models with demography or other basic compartment models, similar expressions for R_0 exists.

If $R_0 > 1$, the introduction of an infected individual in a population has the potential to set off an epidemic. Otherwise, if $R_0 < 1$, an added infectious individual would produce less than one new infection on average, and the infection will die out. In other words, $R_0 = 1$ is the threshold that decides whether an epidemic is possible.

If the population is not completely susceptible, the infected individual in question will sometimes 'try' to infect an individual who is already infected. The *effective reproduction number* ($R_e(t)$) [13] is the average number of new infections caused by one individual introduced into such a population, and can be calculated as

$$R_e(t) = R_0 \cdot s(t), \tag{3.2}$$

where $s(t)$ again is the proportion of susceptibles at time t.

The reproduction numbers R_0 and R_e are the most common reproduction

numbers being used, and they have an important role in infectious disease modelling and assessment. There are however other variants, such as the *actual reproduction number* ($R_a(t)$), which describes the actual spread of an epidemic that has already occurred. R_a is calculated as the average number of secondary cases per case to which the infection has actually been transmitted during the infectious period in a population [13,19]. For the basic SIR model, the expressions for $R_a(t)$ and $R_e(t)$ are identical, but this is generally not the case for more complex models.

Another type of quantity that is related to the growth of an epidemic is the epidemic growth rate. The *intrinsic growth rate* (r) [13] denotes the growth rate at the very beginning of an epidemic outbreak. Here, the growth is exponential, and the intrinsic growth rate corresponds to an exponential growth rate. The intrinsic growth rate r differs from the reproduction numbers in the way of describing epidemic growth. While the reproduction numbers describe the reproduction over an individual's infectious period, r describes the immediate growth of the entire epidemic. For long lasting infections, where disease progression can change, or be uncertain, r might be a more meaningful quantity than reproduction numbers such as R_0. An obvious example of such a long lasting infection is HIV, where the situation for infected individuals has changed when new treatment, such as HAART, has become available.

For the basic SIR model in Figure 1, r is found by

$$r = k - \gamma. \tag{3.3}$$

The growth rate can also be expressed as a *time-dependent growth rate* ($r(t)$) [13]. For the same basic SIR model, $r(t)$ is

$$r(t) = \lambda(t) - \gamma, \tag{3.4}$$

or the difference between the 'birth' rate and 'death' rate of new infections

per unit time.

To read more about reproduction numbers and epidemic growth rates, see **Paper 1** or Diekmann and Heesterbeek [12]. In **Paper 1** we also discuss how to estimate these quantities in more advanced compartments models, while estimation from observed data is discussed in both **Paper 1** and **Paper 2**.

4 Causal inference

Saying something about the relationship between cause and effect for certain variables or events is one of the main aims of natural and social sciences, including medical research. However, the methodology for extracting such causal relationships has been the subject of fierce debate over the years [20]. In statistics, maybe even more than in other fields, there has traditionally been a very cautious attitude towards talking about causation, based on an awareness of the fundamental difference between statistical association and actual causal relationships [2]. This attitude has to some extent gradually changed over the last decades, with what could be called a growing causal movement within statistics. People like J. Pearl, J. M. Robins, D. B. Rubin and others began to develop frameworks for causal thinking in statistics, speaking of and defining concepts like causal effects. However, what statisticians speak of under the causality banner varies. One categorisation is to divide the topics into the following three areas of statistical causality: counterfactual causality, graphical models and predictive causality [2].

The methods discussed in the following sections, and in **Paper 3** and **Paper 4** fall within the first two areas. The topics in this section are within counterfactual causality, while Section 5.1 will touch on the area of graphical models and the concepts of direct and indirect effects.

4.1 Counterfactual causality

The main objective in counterfactual causality is to measure the effect of some intervention, such as medical treatment or another preventive health measure.

Take, for example, the causal effect of treatment with regard to patients' lifetime. Every patient i has two possible responses: one if the patient was given treatment, Y_{1i}, and one if the patient was not given treatment, Y_{0i}. Such responses are also called potential outcomes [20]. The *causal effect* of treatment for patient i is then merely

$$\theta_i = Y_{1i} - Y_{0i}. \tag{4.1}$$

Ideally, this would be the causal effect we are interested in, but obviously both potential outcomes are not possible to observe. One of the outcomes can be observed, the other is *counterfactual* and cannot be observed, simply because it does not exist. This has been called the fundamental problem of causal inference [22], but it does not mean that causal inference is impossible.

The alternative to the causal effect in (4.1) is to look at the average causal effect. If Y_1 and Y_0 are the potential outcome random variables, *the average causal effect*, often referred to as just the causal effect, is [21]

$$E[\theta] = E[Y_1 - Y_0] = E[Y_1] - E[Y_0]. \tag{4.2}$$

The definition of a causal effect in equation (4.2) could also be written using *do*-operator notation, where setting a variable X to a value x is denoted by $do(X = x)$, or $do(x)$ for short. Given two disjoint sets of variables X and Y, where x_1 and x_0 are two distinct realizations of X, the causal effect of X on Y, θ, can be denoted by [20]

$$E[\theta] = E(Y|do(x_1)) - E(Y|do(x_0)). \tag{4.3}$$

The average causal effect from (4.2) or (4.3) is found as the average outcome for patients receiving treatment minus the average outcome for patients *not* receiving treatment. Models for estimating such effects are called *counterfactual models* or *potential-outcome models*. For these estimates to be unbiased, there are two main assumptions which need to hold:

1. *The stable unit treatment value assumption (SUTVA)*: The causal effect for a particular patient does not depend on assignments of other patients and there are no hidden versions of treatment [23].

2. *(Strong) ignorability*: Treatment assignment T is independent of the outcomes given a vector of all the observable covariates X, or $(Y_1, Y_0) \perp\!\!\!\perp T|X$ [21]. In other words: the treatment assignment mechanism can be said to be ignorable, given all observed covariates.

The randomized controlled trial (RCT) has for long been seen as the gold standard in medical research [24]. In RCTs the patients are randomly assigned to different groups, such as a treatment group and a control group. The resulting effect estimate is then the difference between the mean response in the different groups. In this way, ideal RCTs fulfill the ignorability assumption through the random treatment assignment. However, the ideal set-up can be violated for many reasons; perfect control over patients is hard to achieve, randomization into placebo groups is not always ethical, and the presence of randomization itself might influence the participation in the study. In other words, it is not always feasible to create a RCT. HIV treatment is a good example of when random treatment assignment would not be ethical: assigning a patient to placebo would be to deny the patient possible life-saving treatment [20]. Studies of such treatment effects would then have to lean on observational studies.

Observational studies are empirical investigations of treatments, policies or exposures, and their effects. The studies may have the same goal as RCTs, but differ in that the investigator cannot control the treatment

assignment [25]. The main examples of such observational studies used in this thesis are the many large HIV cohort studies, especially the Swiss HIV Cohort Study [8] analysed in **Paper 3** and **Paper 4**.

In an ideal RCT, because of the randomization, we have that association is causation. This is not the case in observational studies. Here, in general association is not causation. However, observational studies are often the only alternative for causal inference, and there exist methods that permit the estimation of causal treatment effects from observational data, under certain assumptions [26].

4.2 Causal inference from observational studies

To say something about the average causal effect of an observed treatment status variable D on an outcome variable Y in observational studies, *the treatment selection mechanism* [21] has to be taken into the model. In other words, the bias due to any lack of randomization has to be adjusted for. For pre-treatment differences, there are two types of confounders creating such selection biases: the observed or measured confounders, and the ones not measured, but suspected to exist, the *unobserved* or *unmeasured confounders*. A goal in causal inference of observational studies is then to adjust for all observed confounders, and to assess and reduce sensitivity to possibly unmeasured confounders. The methodology needed to do this will vary with the complexity of the treatment selection and the data.

If the treatment exposure is fixed, and given that there are no unmeasured baseline covariates or model misspecifications, conventional methods to adjust for confounding by baseline covariates such as stratification, matching and/or regression would give consistent estimators of the causal effect. However, when estimating the causal effect of a time-varying treatment on an outcome, conventional methods could fail to have a causal interpretation, even when there are no unmeasured confounders, no model misspecification, and all of the time-dependent confounders are controlled for [28].

HIV cohort data, such as the Swiss HIV Cohort Data, are a typical example where such problems of time-dependent confounding are present. *Time-dependent confounding* could be present when a variable, affected by past treatment, is both a predictor of future treatment and of outcome [29].

In the HIV cohort setting, patients are typically observed through repeated visits, where at each visit the treatment indicator variable and variables such as the outcome and other relevant covariates are updated. An example of a time-dependent confounder would be the CD4 cell count, which because treatment for HIV is not given before the patient is at a certain stage of disease, is a predictor of future treatment and outcome (usually AIDS or death). Also, treatment would again increase the CD4 count; past treatment will affect later CD4 levels. Methods for estimating the causal effect of treatment in the presence of time-dependent confounding in such datasets will be the theme for the rest of this section.

4.2.1 Marginal structural models

The most established method for estimating the causal effect of a time-dependent exposure A on an outcome Y in the presence of time-dependent confounding covariates L is probably the *marginal structural model* (MSM) [29, 30]. Using the notations of Robins, A_k and L_k denote the values of the time-dependent variables A and L at the discrete time point k, such as in the k^{th} month since start of follow-up.

The models are called *marginal* because they model the marginal distribution of the counterfactual random variables $Y_{a_0=1}$ and $Y_{a_0=0}$, and *structural* because they model the probabilities of counterfactual variables, often called structural models in econometrics and the social sciences [29]. The parameters of the MSM can be estimated using so-called *inverse probability of treatment* (IPT) weights, and implemented as an extension of standard methods such as the Cox proportional hazards model.

One way to understand the MSMs and the IPT weighting procedure is

to think of the problem of unmeasured counterfactuals as a missing data problem. Ideally we would want the 'full data', including all potential outcomes for every individual, but in our observed data the counterfactuals are obviously missing. However, the situation is similar to when dealing with censored data, and a method for adjusting for dependent censoring such as inverse probability weighting can be used.

Consider an observational study, such as a HIV cohort study, with non-random and monotone missingness (if the patient drop out, he does not come back). When dealing with dependent censoring, *inverse probability of censor* (IPC) weights [31,32] are used to account for the missing observations in the dataset. The idea is to 'rebuild' the dataset to what it would have looked like without any censoring, by weighting each observed individual at time t with the inverse probability of being observed at time t. This corresponds to weighting each individual to account for the number of 'similar' individuals who have dropped out. The probability of being observed, or 1 minus the probability of being censored, could be estimated using logistic regression, adjusting for all relevant covariates. In addition, inverse probability weights are often stabilized to achieve greater efficiency (see next paragraph) [30].

Continuing with a notation similar to the one by Robins and Hernan, let $C(t)$ be 1 if a patient is censored at time t and 0 otherwise, V a vector of time-independent baseline covariates, $\bar{A}(t) = \{A(u); 0 \leq u < t\}$ the covariate history up to time t, and similarly for $\bar{C}(t)$ and $\bar{L}(t)$. $\bar{A}(0)$ and $\bar{C}(0)$ is by definition 0. The stabilized IPC weight for patient i at time t is

$$w_i(t) = \prod_{k=0}^{t} \frac{P[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1) = \bar{a}_i(k-1), V = v_i)]}{P[C(k) = 0 | \bar{C}(k-1) = 0, \bar{A}(k-1) = \bar{a}_i(k-1), \bar{L}(k-1) = \bar{l}_i(k-1)]}. \tag{4.4}$$

Using unstabilized weights would correspond to using the weights from equation (4.4), but replacing the numerator with 1. This way of weighting for

dependent censoring using stabilized IPC weights is used in both **Paper 3** and **Paper 4** in this thesis.

A MSM can be fitted in a similar way using IPT weights, weighting for the 'missing' counterfactuals. To simplify, treated patients are weighted with the inverse probability of being treated and untreated individuals are weighted with the inverse probability of being untreated. The IPT weight for patient i at time t can be written

$$w_i^*(t) = \prod_{k=0}^{t} \frac{P[A(k) = a_i(k)|\bar{A}(k-1) = \bar{a}_i(k-1), V = v_i)]}{P[A(k) = a_i(k)|\bar{A}(k-1) = \bar{a}_i(k-1), \bar{L}(k-1) = \bar{l}_i(k-1)]}. \tag{4.5}$$

A MSM could be fitted adjusting for dependent censoring, using the combined weights $w_i(t) \times w_i^*(t)$.

The idea is that, in the weighted dataset, treatment is independent of the confounders. The MSM models a situation where treatment is randomized.

The MSM relies on four main assumptions [33]:

1. *Consistency.* Each individual's counterfactual outcome under their observed exposure is the same as their observed outcome.

2. *Exchangeability.* This breaks down to an assumption of no unmeasured confounders, and is often referred to as the *no unmeasured confounders* assumption.

3. *Positivity.* There must be both exposed and unexposed individuals at every level of the confounders, so that there are positive probabilities at all levels. This assumption is sometimes referred to as the *experimental treatment assignment* (ETA) assumption.

4. *No model misspecification.* The MSM must be correctly specified, including the probability of treatment and censoring models.

The assumption of exchangeability (or no unmeasured confounders) is not testable using observed data, but can be explored with sensitivity analysis [33].

The MSM fitted with IPT weights has both strengths and weaknesses [28], one of the weaknesses being that the IPT weights can be unstable. If the time periods, indexed by k in (4.5), are small, the probability in the denominator of the same equation can be very small, resulting in inordinately large weights and both bias and imprecision [28]. The problem with instability in IPT weights is discussed in **Paper 3**.

4.2.2 Other existing methods

There are alternative methods to the IPT weighting of MSMs for estimating causal effects for time-varying exposures in the presence of time-dependent confounders. These methods will give identical effect estimates if certain requirements are maintained, but in general different estimates can be produced. The estimates will depend on both the causal contrast of interest and the method's robustness [28].

One such method is the *g-computation* [35] of MSMs, also referred to by *g-formula*. The g-computation estimation corresponds to maximum likelihood estimation of MSM parameters [34].

Simplified, if $f(y|do(A = a))$ is the distribution of counterfactual or potential outcomes over all individuals, where A is the set of all possible exposure levels a, the causal effect of A on the outcome Y could be evaluated, given that this distribution was known. However, only the distribution of outcomes for the actually received exposures is observed, and the effect of A on Y would be biased when there exists a time-dependent confounder L. The solution called g-computation is to use the factorization [36]

$$f(y|do(A = a)) = \int_l f(y|do(A = a), L = l) f(l) \mathrm{d}l, \tag{4.6}$$

which corresponds to ordinary standardization when exposure is a fixed base-

line characteristic, but not when exposure is time-dependent [1]. Under the assumptions of no unmeasured confounders and consistency, the term in (4.6) is equal to the observable quantity

$$\int_l f(y|A=a, L=l)f(l)\mathrm{d}l, \tag{4.7}$$

which, in simple cases, can be computed directly [28, 36].

G-computation of MSMs could be an option when the ETA assumptions of the IPT weighted MSMs do not hold, and also if it is easier to predict the outcome given exposure and confounding variables than it is to predict the exposure given the confounding variables. However, the g-computation algorithm is often impractical and requires more complicated, iterative procedures than fitting a MSM using IPT weights [36].

Faced with a choice of modelling the exposure, as in the IPT weighted MSM, and modelling the outcome, as in the g-computation of MSMs, a third alternative is to do both, and combine the two approaches. Such methods are formalized as *double robust methods*, where the 'double robust' term reflects that the method has two ways to discover the correct answer [1]. Such methods are not discussed any further in this thesis.

Another alternative method is the *g-estimation* [37] of so-called *structural nested models* (SNMs). G-estimation is based on the concept of *counterfactual failure times*. The idea is that for every individual i, the counterfactual failure time U_i is the failure time that would have occurred if the individual was unexposed throughout follow-up. It is then assumed that exposure accelerates the failure time by a factor of $\exp(-\psi)$, which is called the *causal survival time ratio*. G-estimation is a method for estimating this parameter ψ [38], and is related to the accelerated life models by Cox and Oakes [1].

Formally, the counterfactual failure time $U_{i,\psi}$ can be derived from the observed failure time T_i by

$$U_{i,\psi} = \int_0^{T_i} \exp(\psi e_i(t))dt, \tag{4.8}$$

where $e_i(t)$ is 1 if individual i is exposed at time t and 0 otherwise. In the presence of time-dependent confounding, ψ cannot be estimated using standard accelerated failure time methods, but it can by using g-estimation. The key assumption is that there are no unmeasured confounders or, in other words, that all variables contributing to whether an individual is exposed at a certain time are measured. If this assumption holds, it means that at each measurement time the exposure is independent of the counterfactual failure time U_i. G-estimation is now the search for the value ψ_0, for which exposure at each measurement time is independent of U_{i,ψ_0}. This could be achieved by fitting a logistic regression model, modelling exposure by $e_i(t)$ with $U_{i,\psi}$ and all confounders as covariates. The confounders are then typically the other covariates at the current time point t, and the values of the other covariates and exposure at previous time points and baseline. A series of such logistic regression models are then fitted, with different values of ψ. The g-estimate of ψ, ψ_0, is the value of ψ giving a Wald statistic at zero (a p-value of 1) for the regression coefficient belonging to the $U_{i,\psi}$ covariate [38]. In the presence of censoring, g-estimation becomes more complex [1].

The methods discussed in this section so far (IPT weighting of MSMs, g-computation of MSMs and g-estimation of SNMs) are referred to as *g-methods* [28], where the 'g' stands for generalized, meaning that g-methods generally can be used to estimate the effects of time-varying treatments [39].

However, methods for adjusting for confounding in observational studies can be classified into two categories: g-methods and stratification-based methods [39]. Stratification-based methods include stratification, restriction and matching, and estimating the association between exposure and outcome in subsets of the population, where the exposed and unexposed are assumed to be exchangeable [39].

One way to view all these methods is that they try to mimic the situation of a randomized controlled trial (RCT). The g-methods mimic a RCT by modelling a situation where the exposure is independent of the confound-

ing covariates, for example by weighting. Stratification-based methods can mimic a RCT in a more direct sense, by constructing subsets of the population where exposed and unexposed can be compared without bias. The method discussed in the next subsection, and in **Paper 3**, is a method based on mimicking RCTs by manipulating the observed data.

4.2.3 The sequential Cox method

We still consider estimating the causal effect of an exposure in the presence of time-dependent confounders using observational data. Say that the exposure variable is a dichotomous treatment variable where, once treatment is initiated, individuals by definition do not return to being untreated. The *sequential Cox method* [40] is based on mimicking a sequence of RCTs, each of which can be analysed using a standard Cox proportional hazards model, and then combining these analyses using composite likelihood inference [41]. The method is presented formally in **Paper 3**.

Each of the mimicked RCTs when using the sequential Cox method are constructed based on individuals starting treatment in a certain time interval k, called the reference interval. If the measurement scale in such data is months since inclusion in the study, each observation month could define such a time interval, but there are also other intervals that could be used.

The individuals initiating treatment in time interval k form the treatment group in the mimicked trial, while the individuals not yet on treatment by interval k form the control group. However, the individuals not yet on treatment by interval k are artificially censored from the mimicked trial if they start treatment at a later time point (in an interval $> k$). Then, in the Cox proportional hazards analysis of one such mimicked trial, baseline covariates and covariate values at time interval k would typically be controlled for. Time-dependent confounding variables would be affected by exposure, and are thus not included. Only covariate values at baseline and at the 'local baseline' in the mimicked trial (covariate values at interval k) are included.

If there is bias due to dependent censoring, this could be adjusted for using IPC weights, along with the dependent censoring introduced by the artificial censoring needed to construct the mimicked trials. Including such weights, the estimated parameter in a weighted Cox model would have a causal interpretation if:

1. there are no unmeasured confounders,
2. the model for estimating the hazard rate is correct, and
3. the model for estimating the censor weight is correct.

The idea is now to construct such mimicked RCTs for every possible reference interval k. Say there are K such trials to mimic. The partial score functions from each of the K mimicked trials can be combined using composite likelihood techniques [41] to estimate an overall effect of treatment. See **Paper 3** for details. This overall estimate will still have a causal interpretation given the assumptions that:

1. the treatment effect is the same in every trial, and
2. the effect of treatment is the same for all covariate histories before the reference interval k, given covariates at interval k.

However, the first of these two assumptions can be relaxed by interpreting the overall effect estimate as an average or aggregated causal effect over all mimicked trials. Using the composite likelihood to combine the analyses is similar to taking a weighted average.

As for the other methods, the sequential Cox method has advantages and disadvantages. One advantage compared to the IPT weighted MSM is that the problem of unstable IPT weights is avoided (see **Paper 3** for further discussion). Another advantage is that possibilities open up due to the conditioning on covariate values at time k; this makes it possible to investigate treatment effects for different covariate values at the time of treatment start.

A third advantage is that the model is easy to implement using standard methods and software once you have constructed a dataset for each of the K mimicked trials. Then the overall effect estimate could be estimated using a stratified weighted Cox model, stratified on k. A disadvantage is that you could end up creating a very large pseudo dataset, consisting of data for all the K constructed trials. Also, the confidence intervals based on the Wald statistics would not be valid, and using bootstrap methods can be time consuming.

5 Direct and indirect effects

The concepts of *direct* and *indirect effects* are connected to a mechanistic view of causality [2]. The idea is that the effect of treatment can be divided into a direct effect and an indirect effect. An indirect effect is the effect of treatment going through some intermediate variable. Depending on the situation, there can be many such indirect effects of one treatment going through different intermediate variables. The direct effect is the remaining effect not mediated through any of these intermediate outcomes.

Path analysis is a method for estimating such direct and indirect effects for a set of variables, using a pre-specified system of causal relationships. These relationships between variables are usually represented in a path diagram, which we will describe further in Section 5.1.2. Path analysis was originally developed by geneticist Sewall Wright in the 1920s [42, 43]. Decades later, in the 1950s and 1960s, it was introduced in economics and sociology. During the 1970s it grew more popular and emerged in numerous papers in fields such as psychology, political science, ecology and others [44], and it was generalized into *structural equation modelling* (SEM).

Since it was introduced, path analysis has been subject to criticism and debate considering the value of the method, especially with regard to causality. When Wright used the word 'causal' in connection with his first path

models, it started a furious debate with Henry Niles [45], even though he urged people to be cautious about the use. Many of the criticisms by Niles, and the responses to them, have been repeated in the modern discussion of path analysis and SEM. Most of the criticism is related to cases where there is insufficient information to specify causal relations and where the usefulness of applying the model is unclear [46].

More recently, in relation with to growing causal movement in statistics, there has been new attempts to handle many of these problems using counterfactual frameworks, structural model frameworks, graphical methods and combinations of these. Authors such as Robins [37, 49], Greenland [48], Pearl [50], Rubin [51], van der Lan [52] and many others have developed methods for the identification of direct and indirect effects [47].

It is also being argued that when the goal is to understand underlying mechanisms, it is natural to think in less definite terms than tends to be the case for statistical causal inference. Here, causality is usually viewed as absolute; either the effect is causal, or it is only an apparent effect due to confounding or bias [2]. Even in cases where the information on causal relations is insufficient and the usefulness of the model unclear, a specified model could still be used to explore relations that will be tested in subsequent studies [46].

When discussing direct and indirect effects in the remainder of this thesis, the focus will mainly be on topics related to the dynamic path analysis of **Paper 4**. However, there will also be given a short introduction to graphical models, such as causal directed acyclic graphs and path diagrams, which are related topics when estimating such effects. To read more on classical path analysis, SEM and direct and indirect effects in general, see the references given above or Pearl [20].

5.1 Graphical models

A *graphical model* is a statistical or probabilistic model where dependencies between random variables are represented by means of a graph [53]. Graphical models in statistics is a huge field, including wide classes of models such as Bayesian networks, Markov networks and others. The topics in this section will be limited to two specific types of graphical models, namely causal directed acyclic graphs and dynamic path diagrams. These types of graphical models are used in **Paper 3** and **Paper 4** in this thesis. For other methods, see for example Lauritzen [54].

5.1.1 Causal directed acyclic graphs

Much of the theory around graphical models in causality was presented by Judea Pearl in his book from 2000 [20]. Most of these models are based on the use of *directed acyclic graphs* (DAGs). Often, statements in the counterfactual or potential outcome framework can be represented concisely using graphs, and Pearl also showed that many fundamental concepts in the two different frameworks were equivalent. It was shown that such graphical models provide a direct and powerful way of thinking about causal systems and that identification strategies could be used to estimate effects within such systems [21]. In the following section there will be a short introduction to some of these concepts, starting with some graph terminology [1, 20].

A *graph* is a set V of *vertices* or *nodes* and a set E of *edges*. The nodes usually represent random variables, while the edges, which are links or the arrows in the graph, are connections between the nodes. An edge between two variables indicates a certain relationship between them. An example of a graph is given in Figure 3.

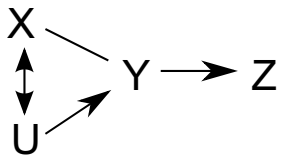

Figure 3: A graph with four nodes X, Y, Z and Y.

An edge could be *undirected*, like the edge (X, Y), *directed*, like the edges (U, Y) and (Y, Z), or *bidirected* like the edge (X, U) [20]. If all edges are directed, the graph is called a *directed graph*. A *path* in a graph is a sequence of edges connected to each other, such as $((Y, X), (X, U), (U, Y), (Y, Z))$ in Figure 3. This is regardless of the direction of the edges. A *directed path* is a special case where all the edges are directed from the starting node to the end node. Any other path is an *undirected path*. If there exists a path between two nodes, these nodes are said to be *connected*, otherwise they are *disconnected*. A *cyclic graph* is a graph where there exists at least one path going out of a node which can be followed through directed edges back to the original node. An *acyclic graph* is a graph that contains no such cycles. So, a directed acyclic graph (DAG), is a graph that is both directed and acyclic, such as the graph in Figure 4.

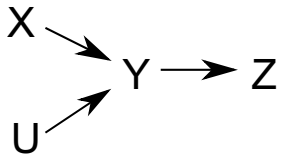

Figure 4: A direct acyclic graph (DAG).

A node, or variable, where two arrowheads meet, such as Y in Figure 4, is called a *collider*. The *children* of a variable are the variables that are affected directly by that variable. For example, Y is a child of both the variables X

and U, and Z is a child of Y. Correspondingly, *parents* of a variable are the variables directly affecting that variable; X and U are the parents of Y. Generally, variables affecting another variable, directly or indirectly through a certain path, are called *ancestors*, while variables being affected, directly or indirectly, are called *descendants*. A node, or variable, without parents is called *exogenous*, otherwise it is *endogenous*.

A causal DAG is a DAG where the arrows can be interpreted as causal relationships and where all common causes of any pair of variables in the graph are included [55].

Generally, as listed in [1], four different causal structures can contribute to the association between two variables X and Y: 1) X causes Y, 2) Y causes X, 3) X and Y share a common cause which has not been conditioned on (confounding), or 4) it has been conditioned or selected on a variable which is affected by X and Y, on a variable which is influenced by such a variable, or on a variable sharing causes with X and Y (collider bias).

Besides giving a visual representation of a causal system, the methodology around causal DAGs is a powerful tool for identifying causal relations and for recognizing and avoiding common mistakes in causal analysis. One such important tool is *d-separation*, or *directed graph separation*. Here, a path is called *open* or *unblocked* if it has no colliders along it. If there is a collider on the path, it is called *closed* or *blocked*. Two variables are said to be *d-separated* if there is not an open path between them, and otherwise they are said to be *d-connected*. If two variables are d-separated, it implies that they are marginally independent [1].

There is also a concept of graphical conditioning, called *conditional d-separation*, which can be summarized by the following rules: 1) to condition on a non-collider Z on a path blocks the path at Z, and 2) to condition on a collider W, or a descendant of W, opens the path at W.

There are also other concepts. An undirected path from X to Y is called a *back-door path* if it starts with an arrow pointing into X. To identify the

causal effect of X on Y, all back-door paths must be blocked. A set of variables $\boldsymbol{S}$ is said to satisfy the *back-door criterion* for identifying the effect of X on Y if it contains no descendants of X and there are no open back-door paths from X to Y when conditioning on $\boldsymbol{S}$.

To read more about these and related methods see Pearl [20] and Rothman, Greenland and Lash [1]. Some of the concepts discussed in this section are used in **Paper 3**.

5.1.2 Path diagrams

In path analysis, or more generally in SEM, a *path diagram* is a graphical representation of the model in question. The diagram is equivalent to a set of equations defining this model (in addition to distributional assumptions) [56]. Unlike the DAGs, dynamic path diagrams could have bidirected arrows, representing correlation between variables. For more on path diagrams for SEM, see for example Raykov and Marcoulides [56] and Spirtes *et al.* [57].

A *dynamic path diagram* is a version of the regular path diagram, which illustrates dependencies for time-dependent variables. It can be viewed as a set of time-indexed DAGs for all $t \in [0, \infty)$ [58]. An example of a dynamic path diagram can be seen in Figure 5 [2]. Here, A is a fixed covariate, for example a treatment indicator, and $L(t)$ is a time-dependent covariate. $dD(t)$ is the increment of some event of interest, measured by some counting variable $D(t)$, for example death. $\alpha_1(t)$, $\beta_1(t)$ and $\beta_2(t)$ are the *path coefficients*, where, for example, $\beta_2(t)$ describes the effect of the covariate L on dD at time t.

The use of dynamic path diagrams is related to the method of dynamic path analysis, which is the topic in the next subsection and in **Paper 4**. For more on dynamic path diagrams, see Fosen *et al.* [58].

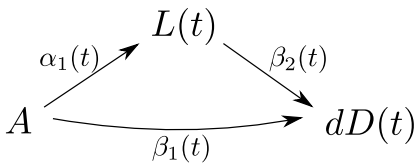

Figure 5: A dynamic path diagram with three nodes, A, $L(t)$ and $dD(t)$, path coefficients $\alpha_1(t)$, $\beta_1(t)$ and $\beta_2(t)$.

5.2 Dynamic path analysis

In *path analysis* the aim is to estimate the path coefficients from a path diagram, analogous to the path coefficients in the dynamic path diagram in Figure 5. The main assumption is that each variable is a linear combination of all the parents in the path diagram. Estimating the path coefficients in a regular path analysis would correspond to fitting a number of linear regression models, regressing each variable or node in the graph onto all of its parents. The effect belonging to a certain path in the graph, from one variable to the outcome variable of interest, such as the indirect effects, is found as the product of all the estimated path coefficients along that path. The total effect can be found as the marginal effect on the outcome, but additivity also provides that the sum of the direct effect and the indirect effects add up to the total effect. To read more about regular path analysis and SEM in general, see for example John C. Loehlin [59].

Dynamic path analysis is a generalization of regular path analysis; it is a combination of path analysis and Aalen's additive regression model [2]. Dynamic path analysis is carried out using recursive least squares regression, as in regular path analysis. Each node in the dynamic path diagram is regressed onto its parent for each time t. The set of variables can be divided into a *covariate set* $(L_1(t), L_2(t), ..., L_p(t))$ and an outcome. In contrast to regular path analysis, the outcome in dynamic path analysis is an *outcome process* $dD(t)$. See the dynamic path diagram in Figure 5. In dynamic path

analysis, the covariates can be both fixed and time-dependent. The path coefficients going into the node of the outcome process are estimated using the additive regression model, with the outcome process as the dependent variable and its parent variables as covariates.

Analogous to regular path analysis, the effect attributable to a certain path at time t in the dynamic path diagram can be calculated as the product of all the path coefficients belonging to that particular path at time t. This is done for all possible times t, and the direct and indirect effects are plotted against time, as the time-dependent effect or regression function itself, or as cumulative effects. Using the example in Figure 5, the direct effect of A is $\beta_1(t)$, while the indirect effect of A is the effect $\alpha_1(t)\beta_2(t)$, going through $L(t)$. Again, the total effect is found as the marginal effect, and also the sum of the direct and the indirect effects.

Dynamic path analysis still has some limitations, discussed for example by Martinussen [60]. The most important reason to be cautious when interpreting estimates of direct and indirect effects is the possible bias due to unmeasured confounders. This poses an even bigger problem when estimating such effects than when estimating marginal or total effects. With the latter, for the estimate to be unbiased, there should be no unmeasured confounders between treatment and the outcome. When estimating direct and indirect effects, there should in addition be no unmeasured confounders between the mediator and the outcome. See Figure 6 for an illustration of such an unmeasured confounder $U(t)$ in the dynamic path diagram from Section 5.1.2. Assumptions of no unmeasured confounders are in general not testable, and will be even more difficult to control in more advanced path models than in the regular regression framework.

Despite the reservations, dynamic path analysis is a useful tool for better understanding how the effect of a treatment on a time-to-event endpoint is mediated through time-dependent covariates [60]. See **Paper 4** for a further discussion on these topics. For more detailed reading on dynamic path

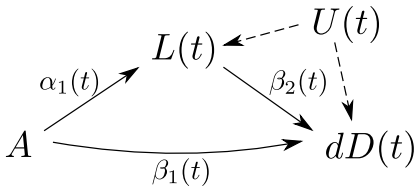

Figure 6: A dynamic path diagram showing a possible unmeasured confounder $U(t)$.

analysis, see especially Fosen *et al.* [58], Aalen *et al.* [2] and Martinussen [60].

5.2.1 Composite dynamic path analysis

In **Paper 4** we discuss a situation where it is reasonable to be interested in direct and indirect effects, or path coefficients for that matter, averaged over many dynamic path analyses. Such analyses could be the dynamic path analyses of multiple mimicked RCTs, such as the ones being constructed using the sequential Cox method in Section 4.2.3 and in **Paper 3**.

Multiple dynamic path analyses can be combined by aggregating over their estimating equations, which corresponds to the composite likelihood inference [41] used in Section 4.2.3 or, more loosely speaking, to a kind of weighted averaging over the different analyses. To estimate aggregated path coefficients, both such composite linear regression models and composite additive regression models have to be fitted. Using artificially censored datasets such as the mimicked randomized trials from Section 4.2.3 and **Paper 3**, dynamic path analysis is generalized into *composite weighted dynamic path analysis* [61]. See **Paper 4** for more details on both aggregating and weighting dynamic path analysis models.

6 Summary of the papers

The current section gives a brief summary of the four papers that form this thesis.

6.1 Paper 1: 'Growth rates in epidemic models: Application to a model for HIV/AIDS progression'

In this paper we discuss the use of epidemic growth rates and reproduction numbers, with application to compartment models for infectious diseases. The compartment model used to illustrate the methods is a generalized SIR model for HIV/AIDS progression, based on a Markov model in an earlier paper by Aalen *et al.* [18].

We discuss the use of these different reproduction numbers and growth rates, such as the basic reproduction number R_0, the effective reproduction number $R_e(t)$, the actual reproduction number $R_a(t)$ and the intrinsic growth rate r. We also discuss the relationship between them, study their behavior using our HIV/AIDS progression model and briefly discuss how these quantities can be estimated from observed data.

The main aim of the paper, beside summarizing the different reproduction numbers and growth rates, is to communicate that, apart from the basic reproduction number R_0 (the predominant quantity), many other measures of epidemic growth will give important supplementary information, and in some situations be more relevant than R_0.

6.2 Paper 2: ' Estimating influenza-related excess mortality and reproduction numbers for seasonal influenza in Norway, 1975–2004'

The main aim of this paper is to estimate the mortality attributable to seasonal influenza in Norway, using surveillance data on influenza-like illness

(ILI) and data on overall mortality. Such data is available from the Norwegian Notification System for Infectious Disease and the Cause of Death Register at the Norwegian Institute of Public Health. Even though most influenza patients make a full recovery, it can be a deadly disease for the elderly and patient groups with certain underlying severe diseases. Therefore, measuring the impact of influenza on mortality is an important public health priority.

In addition to the estimation of excess mortality, we also study the estimation of reproduction numbers. This work can be seen as an addition to the work on reproduction numbers and growth rates in **Paper 1**. In the paper we discuss methodological differences in existing methods for estimating excess mortality, used on data from other countries, before we estimate the excess mortality in Norway from 1975 to 2004 using a Poisson regression based model. Excess mortality is estimated for the overall population and for different age groups.

We estimate overall influenza-related excess mortality in Norway to be an average of 910 deaths per season, or 2.08% of overall deaths. Age-grouped analyses indicate that the major excess mortality is among the elderly in the population. Estimates of the reproduction numbers for the different seasons range from about 1.00 in seasons with no significant outbreak to 1.69.

6.3 Paper 3: 'A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort study'

In this paper we discuss the estimation of the causal effect of treatment from observational data, in the presence of time-dependent confounding. As in **Paper 1**, the focus is on HIV/AIDS, here using data from the Swiss HIV Cohort Study.

We introduce a method based on mimicking a sequence of randomized controlled trials, and analyse them together using Cox regression and composite likelihood inference. We compare our method to the MSM fitted using IPT weights, which is perhaps the most established method for controlling for such time-dependent confounding.

We show how the sequential Cox method introduced in this paper avoids the use of IPT weights and the instability problems associated with the use of such weights in a MSM. We discuss the differences between the two models, the parameter being estimated, and how these differences lead to different possibilities when it comes to estimation. We also see that when estimating the overall effect of HIV treatment on data from the Swiss HIV Cohort Study, both the sequential Cox method and the IPT fitted MSM give similar results.

6.4 Paper 4: 'Analysing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study'

In this paper the aim is to estimate direct and indirect effects of treatment for HIV using dynamic path analysis. As in **Paper 3**, the application is to data from the Swiss HIV Cohort Study.

As in the previous paper, we construct and analyse mimicked randomized trials from the original data. We extend regular dynamic path analysis to include censor weighting and define composite estimates of direct and indirect effects, combining all the mimicked trials.

We show that even simple dynamic path models, with one mediating variable such as the HIV-1 RNA or CD4 level, serve as a useful tool to investigate underlying processes that are often ignored when estimating total treatment effects. The results when applying these models to the Swiss data show that most of the treatment effect goes through the RNA level for the first three or four years after starting treatment. Using CD4 instead of RNA

as the intermediate variable results in a smaller estimate of the indirect effect. We also discuss the limitations of such methods, and the need for caution when interpreting estimates of direct and indirect effects, for example with regard to possibly unmeasured confounders.

7 Discussion

The work in this thesis covers two main areas: infectious disease modelling and causal inference. This division is most apparent between the first two and last two papers. However, setting this distinction aside, there are many other aspects binding these four papers together, some of which were mentioned in the introduction. Common to all papers is that they deal with infectious diseases. **Paper 1**, **Paper 3** and **Paper 4** are all applied in a HIV/AIDS setting, while **Paper 2** is applied to influenza data. Studies of infectious disease are closely linked with understanding mechanisms. Statements therefore tend to be more causal than in studies of chronic diseases. For example, when estimating the effect of some countermeasure that modifies the transmission rates in a multi-stage model such as the one in **Paper 1**, the goal is in fact to estimate the causal effect of that countermeasure, similar to the causal effect of treatment estimated in **Paper 3**. Even in **Paper 2**, when using the influenza level in one week to model the number of deaths a week later, the fact that one event precedes the other indicates that a certain mechanism is assumed. The similar way of thinking about mechanisms in infectious diseases and in causality is also evident through the use of compartment or network models in infectious disease modelling and path models and graphs in causal inference. Consider for example the compartment model in **Paper 1** and the dynamic path models in **Paper 4**.

The main contribution of **Paper 1** is the summary of the different growth rates and reproduction numbers for describing epidemic growth, and the study of how these growth rates behave and relate to each other. Another

contribution is the model for HIV/AIDS disease progression, as a generalized SIR model and an extension of previously used Markov models. Possible future extensions of this work would be to adjust the model to the situation now being seen after the introduction of HAART treatment. Early attempts on this can be seen in Aalen *et al.* [14], but that was just three years after HAART was introduced. Also, in connection to causal inference, a future extension could be to study the use of this kind of transition models in estimating treatment effects.

As for **Paper 2**, the main contribution is the estimates of influenza-related excess mortality in Norway between 1975 and 2004, both overall and within age groups. Estimates of such numbers are of interest to public health representatives, such as the Norwegian Institute of Public Health, and to others working with decision-making and in related areas. Another contribution in this paper is the estimation of growth rates and reproduction numbers for the same data, as few estimates from seasonal influenza are available in the literature. Possible extensions to the work in this paper would be to compare the methods used with other methods for estimating excess mortality, and to make more detailed models based on geographical data, which could make it easier to include other relevant covariates such as temperature. Data for the years between 2004 and 2010 are also now available to be studied.

The main contribution of **Paper 3** is the introduction of a new approach for estimating causal effects of treatment from observational studies, combining the ideas of mimicking randomized trials and composite likelihood inference. Using data from the Swiss HIV Cohort Study, we show results similar to the ones using a MSM, and that the method can be implemented using existing methods when a manipulated version of the original dataset is created. Possible extensions to this work would be to study the method more systematically, for example using simulation studies.

In **Paper 4**, the main contribution is the extension of dynamic path

analysis, introducing both weighting and composite analysis, to analyze the same type of observational data as used in the previous paper. Possible future work would be to study how and when models with more complex covariate dependencies can be used, and also to address the remaining challenges in formalizing the method of dynamic path analysis in general.

References

1. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. Lippincott Williams & Wilkins, 2008.

2. Aalen OO, Borgan Ø, Gjessing HK. *Survival and Event History Analysis: A Process Point of View*. Springer: New York, 2008.

3. Anderson RM, May RM. *Infectious diseases of humans – Dynamics and control*. Oxford University Press, 1991.

4. Keeling MJ, Rohani P. *Modeling Infectious Disease in Humans and Animals*. Princeton University Press, 2008.

5. Sterne JAC, Hernan MA, Ledergerber B, Tilling K, Weber R, Sendi P, Rickenbach M, Robins JM, Egger M, Swiss HIV Cohort Study. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet* 2005; **366**:378–384. DOI: 10.1016\S0140-6736(05)67022-5

6. Volberding P, Sande MA. *Global HIV/AIDS medicine*. Elsevier Health Sciences, 2008.

7. Kaslow RA, Ostrow DG, Detels R, Phair JP, Polk BF, Rinaldo CR, Jr. The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* 1987; **126**(2):310–318.

8. Ledergerber B, Egger M, Opravil M, Telenti A, Hirschel B, Battegay M, Vernazza P, Sudre P, Flepp M, Furrer H, Francioli P, Weber R. Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. Swiss HIV Cohort Study. *Lancet*. 1999; **353**(9156):863–868. DOI: 10.1016\S0140-6736(99)01122-8

9. Bridges CB, Kuehnert MJ, Hall CB. Transmission of Influenza: Implications for Control in Health Care Settings. *Clinical Infectious Diseases* 2003; **37**(8):1094-1101. DOI: 10.1086/378292

10. Monto AS, Gravenstein S, Elliott M, Colopy M, Schweinle Jo. Clinical Signs and Symptoms Predicting Influenza Infection. *Archives of Internal Medicine* 2000; **160**(21):3243–3247. DOI: 10.1001/archinte.160.21.3243

11. Gran JM, Iversen B, Hungnes O, Aalen OO. Estimating excess mortality from influenza in Norway 1975–2004. *Epidemiology and Infection* 2010; **138**(11):1559–1568. DOI: 10.1017/S0950268810000671

12. Diekmann O, Heesterbeek JAP. *Mathematical Epidemiology of Infectious Disease – Model Building, Analysis and Interprettation*. Wiley, 2000.

13. Gran JM, Wasmuth L, Amundsen EJ, Lindqvist BH, Aalen OO. Growth rates in epidemic models: Application to a model for HIV/AIDS progression. *Statistics in Medicine* 2008; **27**(23):4817–4834. DOI: 10.1002/sim.3219

14. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. New therapy explains the fall in AIDS incidence with a substantial rise in number of persons on treatment expected. *AIDS* 1999; **13**(1):103-108. DOI: 10.1097/00002030-199901140-00014

15. Fouillet A, Rey G, Wagner V, Laaidi K, Empereur-Bissonnet P, Le Tertre, A, Frayssinet P, Bessemoulin P, Laurent F, De Crouy-Chanel P, Jougla E; Hémon D. Has the Impact of Heat Waves on Mortality Changed in France Since the European Heat Wave of Summer 2003? *International Journal of Epidemiology* 2008; **7**(2):309–317. DOI: 10.1093/ije/dym253

16. Wong CM, Yang L, Chan KP, Leung GM, Chan KH, Guan Y, Lam TH, Hedley AJ, Peiris JSM. Influenza-Associated Hospitalization in a Subtropical City. *PLOS Medicine* 2006; **3**(4):485–492. DOI: 10.1371/journal.pmed.0030121

17. Nunes B, Natário I, Carvalho ML. Time series methods for obtaining excess mortality attributable to influenza epidemics. *Statistical Methods in Medical Research* 2010;1–15. DOI:10.1177/0962280209340201

18. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. New therapy explains the fall in AIDS incidence with a substantial rise in number of persons on treatment expected. *AIDS* 1999; **13**(1):103–108.

19. Amundsen EJ, Stigum H, Røttingen JA, Aalen OO. Definition and estimation of an actual reproduction number describing past infectious disease transmission: application to HIV epidemics among homosexual men in Denmark, Norway and Sweden. *Epidemiology and Infection* 2004; **132**(6):1139–1149. DOI: 10.1017/S0950268804002997

20. Judea Pearl. *Causality – Models, Reasoning, and Inference*. Cambridge University Press, 2009.

21. Morgan SL, Winship C. *Counterfactuals and Causal Inference – Methods and Principles for Social Research*. Cambridge University Press, 2009.

22. Holland P. Statistics and Causal Inference. *Journal of the American Statistical Associations* 1986, **81**(396): 947.

23. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health* 2000, **21**:121–145. DOI: 10.1146/annurev.publhealth.21.1.121

24. Meldrum ML. A brief history of the randomized controlled trial. From oranges and lemons to the gold standard. *Hematology/Oncology Clinics of North America* 2000, **14**(4):745–760, vii. DOI: 10.1016/S0889-8588(05)70309-9

25. Rosenbaum PR. *Observational Studies*. Springer: New York, 2002.

26. Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *Journal of Epidemiology and Community Health* 2007; **60**(7):578–586. DOI: 10.1136/jech.2004.029496

27. Rosenbaum PR. Observational Study. *Encyclopedia of Statistics in Behavioral science*. John Wiley & Sons: Chichester, 2005.

28. Robins JM, Hernan MA. Estimation of the causal effects of time-varying exposures. *Longitudinal Data Analysis*. Chapman & Hall, 2009.

29. Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 2000; **11**(5):550–560. DOI: 10.1097/00001648-200009000-00011

30. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. *Epidemiology* 2000; **11**(5):561–570. DOI: 10.1093/aje/kwi216

31. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 1952; **47**(260):663–685.

32. Robins JM, Rotnitzky A, Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* 1995; **90**(429): 106–121.

33. Cole SA, Hernan MA. Constructing Inverse Probability Weights for Marginal Structural Models. *American Journal of Epidemiology* 2008; **168**(6):656–664. DOI: 10.1093/aje/kwn164

34. Neugebauer R, van der Laan MJ. G-computation estimation for causal inference with complex longitudinal data. *Computational Statistics & Data Analysis* 2006, **51**:1676–1697.

35. Robins JM. A new approach to causal inference in mortality studies with sustained exposure periods – application to control of the healthy worker survivor effect. *Mathematical modelling* 1986; **7**:1393–1512.

36. van der Wal WM, Prins M, Lumbreras B, Geskus RB. A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. *Statistics in Medicine* 2009; **28**:2325–2337. DOI: 10.1002/sim.3629

37. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. *Epidemiology* 1992; **3**:319–336.

38. Sterne JAC, Tilling K. G-estimation of causal effects, allowing for time-varying confounding. *The Stata Journal* 2002; **2**(2):164–182.

39. Hernan MA, Robins JM. *Causal inference without models*. Chapman & Hall/textbackslashCRC, 2011. Draft v1.10.5, available at www.hsph.harvard.edu/faculty/miguel-hernan/causal-inference-book

40. Gran JM, Røysland K, Wolbers M, Didelez V, Sterne J, Ledergerber B, Furrer H, von Wyl V, Aalen OO. A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine* 2010; 1–12. DOI: 10.1002/sim.4048

41. Lindsay B. Composite likelihood methods. *Statistical Inference from Stochastic Processes* 1988, Ed. Prabhu NU. Providence, RI: American Mathematical Society.

42. Wright S. Correlation and causation. *Journal of Agricultural Research* 1921; **20**:557–585.

43. Wright S. The method of path coefficients. *The Annals of Mathematical Statistics* 1934; **5**:161–215.

44. Lleras C. Path analysis. *Encyclopedia of Social Measurement*. Elsevier, 2005.

45. Denis DJ, Legerski J. Causal modeling and the origins of path analysis. *Theory & Science* 2006; **7**(1).

46. MacKinnon DP. *Introduction to statistical mediation analysis*. CRC Press, 2008.

47. Geneletti S. Identifying direct and indirect effects in a non-counterfactual framework. *Journal of the Royal Statistical Society* 2007; **69**(2):199–215. DOI: 10.1111/j.1467-9868.2007.00584.x

48. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology* 1992; **3**:143–155.

49. Robins JM. Semantics of causal DAG models and the identification of direct and indirect effects. *Highly Structured Stochastic Systems*. Oxford University Press, 2003;70–81.

50. Pearl J. Direct and indirect effects. *Proc. 17th Conf. Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2001.

51. Rubin D. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics* 2004; **31**:161–170. DOI: 10.1111/j.1467-9469.2004.02-123.x

52. van der Laan MJ, Petersen ML. Estimation of direct and indirect causal effects in longitudinal studies. *U.C. Berkeley Division of Biostatistics Working Paper Series* 2004.

53. Madigan D, York J, Allard D. Bayesian graphical models for discrete data. *International Statistical Review* 1995; **63**(2):215–232.

54. Lauritzen SL. *Graphical models*. Oxford University Press, 1996.

55. VanderWeele TJ, Hernan MA, Robins JM. Causal directed acyclic graphs and the direction of unmeasured confounding bias. *Epidemiology* 2008; **19**(5):720–728. DOI: 10.1097/EDE.0b013e3181810e29

56. Raykov T. Marcoulides GA. *A first course in structural equation modeling*. Routledge, 2006.

57. Spirtes P, Richardson T, Meek C, Scheines R, Glymour C. Using path diagrams as a structural equation modeling tool. *Sociological Methods & Research* 1998; **27**(2):182–225. DOI: 10.1177/0049124198027002003

58. Fosen J, Ferkingstad E, Borgan Ø, Aalen OO. Dynamic path analysis – a new approach to analyzing time-dependent covariates. *Lifetime Data Analysis* 2006; **12**:143–167. DOI: 10.1007/s10985-006-9004-2

59. Loehlin JC. *Latent variable models: an introduction to factor, path, and structural equation analysis*. Routledge, 2004.

60. Martinussen T. Dynamic path analysis for event time data: large sample properties and inference. *Lifetime Data Analysis* 2010; **16**:85–101. DOI: 10.1007/s10985-009-9128-2

61. Røysland K, Gran JM, Ledergerber B, von Wyl V, Young J, Martinussen T, Aalen OO. Analysing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study. *Manuscript* 2010; 1–22.

1

2

Epidemiol. Infect. (2010), **138**, 1559–1568. © Cambridge University Press 2010
doi:10.1017/S0950268810000671

Estimating influenza-related excess mortality and reproduction numbers for seasonal influenza in Norway, 1975–2004

J. M. GRAN[1]*, B. IVERSEN[2], O. HUNGNES[3] AND O. O. AALEN[1]

[1] *Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway*
[2] *Department of Infectious Disease Epidemiology, Norwegian Institute of Public Health, Oslo, Norway*
[3] *Department of Virology, Norwegian Institute of Public Health, Oslo, Norway*

(*Accepted 24 February 2010; first published online 25 March 2010*)

SUMMARY

Influenza can be a serious, sometimes deadly, disease, especially for people in high-risk groups such as the elderly and patients with underlying, severe disease. In this paper we estimated the influenza-related excess mortality in Norway for 1975–2004, comparing it with dominant virus types and estimates of the reproduction number. Analysis was done using Poisson regression, explaining the weekly all-cause mortality by rates of reported influenza-like illness, together with markers for seasonal and year-to-year variation. The estimated excess mortality was the difference between the observed and predicted mortality, removing the influenza contribution from the prediction. We estimated the overall influenza-related excess mortality as 910 deaths per season, or 2·08% of the overall deaths. Age-grouped analyses indicated that the major part of the excess mortality occurred in the $\geqslant 65$ years age group, but that there was also a significant contribution to mortality in the 0–4 years age group. Estimates of the reproduction number R, ranged from about 1 to 1·69.

Key words: Excess mortality, influenza (seasonal), reproduction numbers.

INTRODUCTION

Influenza is an infection of the respiratory tract caused by the influenza viruses; RNA viruses belonging to the family Orthomyxoviridae [1]. The disease is characterized by acute-onset fever, headache, myalgia, prostration, coryza, and a dry cough, and is usually self-limiting with recovery in 2–7 days [2, 3]. Primary viral, or secondary bacterial pneumonias are common complications of influenza. Most influenza patients recover without sequelae. Mortality is highest in the elderly and in patient groups with certain underlying, severe diseases [1–3]. For these risk groups annual influenza immunization is recommended in many countries.

In the northern hemisphere the virus usually causes annual outbreaks of varying length and severity during the winter seasons. When a new virus variant emerges to which no one is immune, larger epidemics – know as pandemics – may ensue. Last century three worldwide pandemics occurred; in 1918–1919 (Spanish flu), 1957–1958 (Asian flu), and 1968–1970 (Hong Kong flu) [2, 4].

Most developed countries have some sort of surveillance system for influenza measuring the number of patients seeking healthcare, virologically confirmed cases or some other marker for influenza-like illness (ILI) such as absence from school or work. No

* Author for correspondence: J. M. Gran. M.Sc., Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, PO Box 1122 Blindern, 0317 Oslo, Norway.
(Email: j.m.gran@medisin.uio.no)

country has routine surveillance of influenza-related deaths.

It is difficult to discern if the cause of death is from influenza in patients with several other serious underlying illnesses. Influenza is rarely recorded as the cause of death on death certificates in Norway. Internationally many studies have been performed to estimate the mortality due to influenza [5–15], in Norway this has only been done for influenza pandemics [16].

The aim of this work was to estimate the excess mortality due to influenza in Norway by studying the relationship between the number of reported deaths and clinical influenza activity; and to compare these results with information on predominant influenza viruses and estimates of the reproduction number, R, for the different seasons.

METHODS

Data material

Information on clinical ILI was derived from The Norwegian Notification System for Infectious Diseases (MSIS). For the period 1975–1998 all general practices in primary healthcare and outpatient emergency clinics were obliged to report ILI along with other clinical diagnoses on a weekly basis to the Norwegian Institute of Public Health (NIPH). The average reporting coverage was about 60% of around 2000 practices. Rates of ILI were calculated as a proportion of the total population, without adjusting for the reporting coverage.

From autumn 1998 NIPH designated 201 sentinel reporting units based on geographical location, population size and previous reporting frequencies from the 2000 practices mentioned above. These formed about 10% of the practices but about 25% of the reported volume of ILI. The sentinels reported weekly, from week 40 in autumn to week 20 in spring, the number of ILI [using the case definition of 'R80 Influenza' from the International Classification of Primary Care (ICPC)]. The number of consultations and, from 2004–2005 when the information was available, the number of patients on the patient list of general practitioners, were used as denominators.

The weekly recording period was from Friday to Thursday, after which the report card was completed and sent to NIPH by post and entered into a database in EpiInfo 6·04d (CDC, USA). Quality checks for inconsistencies and improbable figures were performed.

Data on all-cause mortality per week by age group were derived from the Cause of Death Register at NIPH. The week number was obtained by defining week 1 as starting on 1 January each year, week 2 on 8 January and so on. Hence there may be a small discrepancy between the week numbering of ILI and deaths.

Information about dominant virus types and variants for each season was obtained from the virological influenza surveillance records in the Department of Virology, NIPH.

Estimating excess mortality

The number of overall deaths per week was modelled using a Poisson regression model, where the mortality rate was explained by the reported number of ILI cases, the week number and the season. For the ILI covariate, we considered different lagging, before choosing the type of lag which gave the best model fit (i.e. explained the most variability). The week number covariate, a factor numbered between 1 and 52, modelled the seasonality in the data, while the season covariate was a substitute for calendar year. Season was used instead of calendar year because a normal influenza season goes from autumn of one year to spring of the next.

To account for change in population size, the Norwegian population size at the beginning of each calendar year was used as an offset. The model is written as:

$$\hat{D}_{j,k} = \exp\left(\log\left(\text{Population}_{j,k}\right) + \beta_{\mathbf{0}} + \beta_{\text{ILI}} \times ILI_{j,k} + \beta_{\text{week}_j} + \beta_{\text{season}_k}\right),$$

where $\hat{D}_{j,k}$ is the predicted number of overall deaths in week number j and season k, $j \in \{1, 2, \ldots, 52\}$, and $k \in \{\text{spring } 1975, 1975/1976, 1976/1977, \ldots, 2003/2004, \text{autumn } 2004\}$. $\text{Population}_{j,k}$ is the total population size of Norway on 1 January for the year containing week j in season k, $\beta_{\mathbf{0}}$ is the intercept coefficient, β_{ILI} the coefficient to the ILI contribution ($\text{ILI}_{j,k}$ being the reported number of ILI cases in week j in season k), β_{week_j} the coefficient for the factor variable week at week j, and finally β_{season_k}, the coefficient associated with season k. To account for any extra Poisson variation, a dispersion parameter was added, making the model a quasi-Poisson model. The analysis used the GLM package in the open source statistical software R version 2.7.0 [17].

The model was fitted separately for the two datasets, one going from 1975 to 1998, the other from 1998

to 2004. The only weeks included in the analysis were the observed weeks and weeks where the ILI levels with the chosen lag were available. For the data from 1998 to 2004, no ILI figures were collected off-season, i.e. between week 20 and week 40.

The excess mortality was estimated by first considering the overall mortality once the influenza contribution had been removed. Rather than setting the influenza contribution to zero in this estimation, leaving the other parameters and covariates as they were, the influenza contribution was set to some threshold value accounting for the ever-present baseline of ILI cases, also observed off-season. As previously mentioned, off-season ILI numbers were not collected in the new reporting system, and one would expect these weeks (if measured) to represent the lowest ILI counts. We therefore removed the corresponding number of high ILI values, and then most of the outbreaks, to find this threshold. In other words, the threshold value was set to the mean of the remaining ILI values, excluding the 20 lowest and 20 highest values for each season. We expected this to give a conservative estimate of the off-season ILI level. The estimated excess mortality related to ILI was then calculated as the difference between the observed mortality and the predicted mortality leaving out the influenza contribution above this threshold.

Estimating the reproduction number

The reproduction number R, also known as the effective reproduction number, denotes the number of secondary infections caused on average by one infected individual being introduced into a population. R relates to R_0, the reproduction number for an individual introduced into a population of only susceptibles. R_0 serves as a threshold value for epidemic growth, where $R_0 = 1$ is the threshold deciding whether an epidemic is possible or not [18, 19]. When a fraction p of a population is protected from infection, the relationship between R and R_0 is given by $R = (1-p)R_0$ [20, 21].

The reproduction number can be estimated through the initial growth rate of an epidemic r [18, 19, 21], which describes the growth of the epidemic in its initial phase, rather than the individual reproduction as with R.

When the initial growth of an epidemic is assumed to be exponential, r can be estimated by fitting a straight line to the natural logarithm of the number of infected individuals at each time point in the initial phase. The slope of this line is then the estimated initial growth rate r.

To determine the initial phase of an epidemic is a challenge, but one method is to consider the goodness of fit [20, 22]. We determined the initial phase by finding the starting point and the endpoint separately. The starting point was found by looking for structural change when modelling the natural logarithm of the number of infected individuals with linear regression, using data from the start of each season until the time of the influenza peak. If there was an influenza outbreak during this period, a breakpoint on the log scale of the ILI curve should exist. The breakpoint was found using the methods described in Zeileis *et al*. [23] (implemented in the R package STRUCCHANGE). These methods test the null hypothesis of no change in regression parameters before and after each possible time point, and then choose the time point giving the lowest P value as the breakpoint, if this is significant. The endpoint, restricted to be at least three points later than the start point and no later than the peak, was chosen by the best goodness of fit in terms of R^2.

Using the estimate of r we were able, by assuming a simple multistage model for the disease spread, to derive an expression for the reproduction number R [21, 24, 25]. Assuming a SEIR model (susceptible, exposed, infectious, recovered) for the spread of influenza, we obtained

$$R = 1 + \frac{r^2 + (k+\gamma)r}{k\gamma},$$

where k^{-1} is the incubation time and γ^{-1} the infectious period [24]. When calculating the R estimates for seasonal influenza we assumed a 2-day incubation time and a 4-day infectious period [20, 26]. A 95% confidence interval (CI) for the estimate of R was found using the lower and upper confidence limits for the estimate of r (derived from the linear regression model), and the formula for R above.

RESULTS

In the Poisson regression analysis, applying a lag of 1 week to the ILI variable gave the best model fit, when also adjusting for week and season variables. Figure 1 shows an overview of the data on ILI and all-cause mortality as used in the analysis.

The estimated excess mortality for season 1975/1976 to season 2003/2004 varied from 217 deaths (5·31/100 000 population or 0·53% of all deaths) in the 1976/1977 season, to 1802 deaths (41·45/100 000

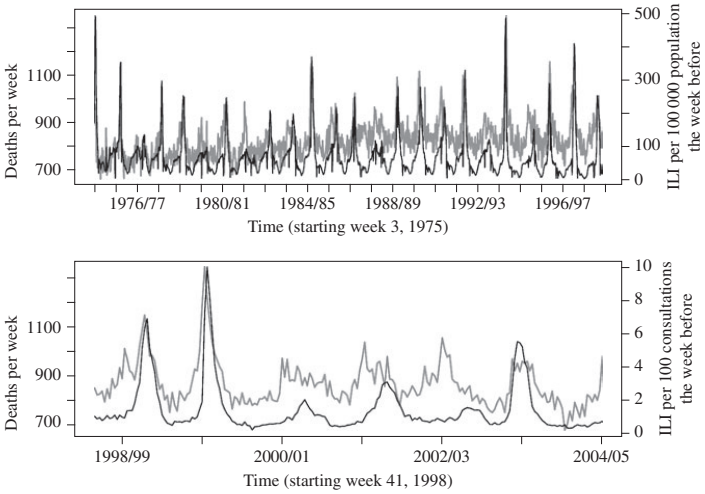

Fig. 1. Overall deaths per week (grey line) and reported number of influenza-like illness (ILI) cases the week before (black line), from the old (upper panel) and new (lower panel) reporting system.

population or 3·89% of all deaths) in the 1993/1994 season: this gave a mean estimated excess mortality of 910 deaths per season (21·25/100 000 population or 2·08% of all deaths), ignoring the two incomplete seasons of spring 1975 and autumn 2004.

Figure 2 shows the observed, overall deaths per week, the overall deaths per week modelled by Poisson regression, and the predicted mortality leaving the influenza contribution fixed at the defined threshold value, under both the old and new reporting systems. The threshold value was found to be 45·91/100 000 population in the data from the old reporting system, and 0·54/100 consultations for the data from the new reporting system. The agreement between the observed mortality and the predicted mortality suggests a good model fit. The estimated dispersion parameters in the analysis of the two datasets were 1·85 and 2·23, respectively. In Figure 2 the estimated excess mortality is the area between the line of the predicted mortality, where the influenza contribution is limited to the threshold value, and the line of the observed mortality (this is most visible in the lower panel).

The estimated excess mortality for each influenza season from 1975/1976 to 2003/2004 is listed in Table 1. This table also gives the total number of ILI cases/100 000 population for the old dataset, the mean number of ILI cases/100 consultations for the new dataset, estimates of R with 95% CIs and dominant virus type for each season. The R estimates range from close to 1, implying no outbreak, to 1·69 in the 1999/2000 season.

In Figure 3 we plotted the estimated excess mortality against estimates of R for each season, marked by groups of dominant virus type. It can be seen that influenza B seasons tended to have lower estimated excess mortality and R estimates than H3N2 seasons. For the other seasons, it can be seen that the two H1N1 seasons both had low estimates of excess mortality and R. Seasons with more than one dominant virus had varying estimates of excess mortality and R, while seasons with no dominant virus had low estimates. The mean estimated excess mortality in H3N2 seasons was 1169 deaths per season, compared to a mean estimated excess mortality of 781 deaths per season in B seasons. The mean estimated excess mortality in H1N1 seasons was 425 deaths per season, but there were only two seasons with H1N1 as the single dominant virus type. When testing the mean difference between estimated excess mortality in H3N2 and B seasons, using a standard two-sample t test, we found that the difference was significant at a 5% level (P value = 0·01). When we did a similar test on the mean difference between estimated R in H3N2 and B seasons, again we found a significant difference (P value = 0·00002).

Table 2 shows estimates of β_{ILI} together with estimated excess mortality, for overall and age-grouped analyses. For the dataset going from 1975 to 1998 the level of ILI had a significant effect on the overall

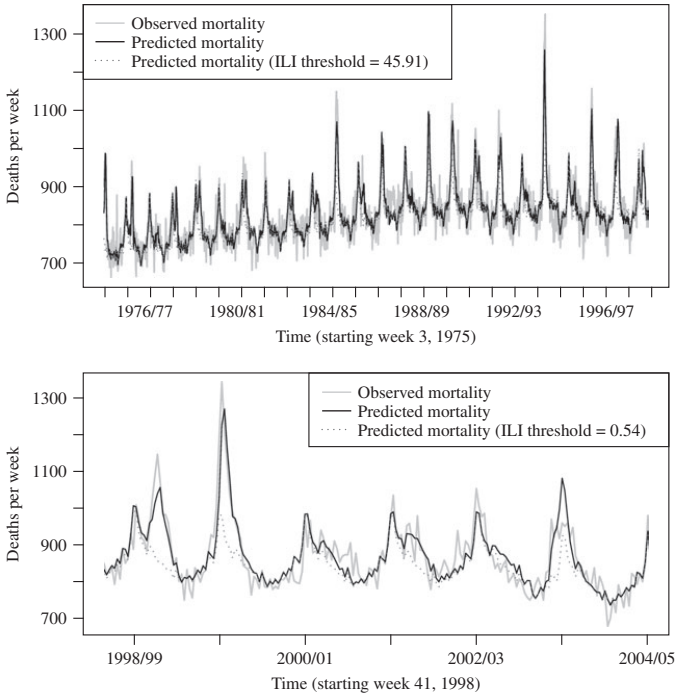

Fig. 2. Observed overall deaths per week (grey line), predicted overall deaths (black line), and predicted mortality where the influenza-like illness (ILI) contribution is limited to a threshold value (dark grey dotted line), for the data from the old (upper panel) and new (lower panel) reporting system. The threshold value accounts for the ever-present baseline of ILI cases, also observed off-season.

deaths in all ages except 5–14 years (using a 5% level of significance). It can also be seen that the highest (significant) effect is in the $\geqslant 65$ years age group, followed by the 0–4 and 15–64 years age groups. Most of the estimated excess mortality was found in the $\geqslant 65$ years age group. Considering the data going from 1998 to 2004 we saw the same patterns, but here the level of ILI had no significant effect on the overall mortality in either the 5–14 or 15–64 years age groups. The effect was higher in the 0–4 years group than in the $\geqslant 65$ years group, but almost all the excess mortality was in the oldest group. Note that the estimated β_{ILI} values for 1975–1998 and 1998–2004 cannot be compared to each other because of the different units for ILI level in the two periods.

DISCUSSION

Our model seems to give a good description of the observed mortality, capturing both the season-to-season change in the number of ILI cases, and the severity. This is also reflected when investigating estimates of the reproduction number R and dominant virus types for the different seasons. The H3N2 seasons had a significantly higher excess mortality and R estimates than the influenza B seasons. For the other strains, there were not many seasons in each group, but the trend was as expected. Seasons with H1N1 and no dominant virus had low impact in terms of excess mortality and R estimates, and the impact varied when there was more than one dominant virus.

ILI data divided into age groups were only available from the 2001 season and onwards, and have not been used in the analysis. Hiwever, age-grouped data for overall deaths were available for all years from 1975 to 2004. The analysis of these data showed that the effect of the influenza outbreaks on overall mortality was highest in the two lowest age groups, and in the $\geqslant 65$ years group. Most of the excess mortality was attributed to the $\geqslant 65$ years group.

Estimation of influenza-related excess mortality has been done in other countries using similar Poisson

Table 1. *Estimated excess mortality for each influenza season, together with numbers of reported ILI cases, estimated reproduction number R and dominant virus type*

Season	Est. excess mortality	Total ILI per 100 000	$\hat{R}$ (95% CI)	Dominant virus
1975/1976	1251	22 076	1·30 (1·17–1·43)	A/Victoria/3/75 (H3N2)
1976/1977	347	22 367	1·35 (0·55–2·46)	A/Victoria/3/75 (H3N2)
1977/1978	651	20 061	1·32 (1·05–1·62)	A/USSR/90/77 (H1N1) and A/Texas/1/77 (H3N2)
1978/1979	912	18 022	1·22 (1·00–1–46)	B/Singapore/222/79
1979/1980	217	23 087	1·02* (1·00–1·04)	No dominant virus
1980/1981	1036	18 696	1·36 (1·25–1·47)	A/Bangkok/1/79 (H3N2)
1981/1982	573	24 174	1·00* (0·06–2·71)	B/Singapore/222/79
1982/1983	422	21 532	1·20 (1·13–1·26)	Sporadic outbreaks of H3N2 and H1N1
1983/1984	831	19 721	1·09 (1·05–1·12)	B/USSR/100/83
1984/1985	1605	18 798	1·31 (1·00–1·65)	A/Victoria/6/84 (H3N2)
1985/1986	1016	17 359	1·08 (1·04–1·11)	B/Ann Arbor/1/86
1986/1987	510	18 089	1·20 (1·09–1·31)	A/Singapore/6/86 (H1N1)
1987/1988	666	22 040	1·07 (1·01–1·13)	B/Victoria/2/87
1988/1989	909	18 616	1·31 (1·10–1·54)	A/Sichuan/2/87 (H3N2)
1989/1990	1291	17 969	1·21 (0·97–1·47)	A/Sichuan/2/87 (H3N2) and B/Victoria/2/87
1990/1991	1044	15 956	1·10 (1·05–1·14)	B/Yamagata/16/88
1991/1992	1303	19 829	1·35 (1·03–1·71)	A/England/261/91 (H3N2)
1992/1993	589	19 914	1·06 (1·02–1·09)	B/Panama/45/90
1993/1994	1802	15 230	1·42 (1·22–1·63)	A/Hong Kong/23/92 (H3N2)
1994/1995	621	18 734	1·15 (1·14–1·17)	B/Beijing/184/93
1995/1996	1077	16 012	1·36 (1·22–1·50)	A/Johannesburg/33/94 (H3N2)
1996/1997	1426	15 746	1·51 (0·91–2·24)	A/Wuhan/359/95 (H3N2) and B/Beijing/184/93
1997/1998	859	14 073	1·45 (1·30–1·60)	A/Wuhan/359/95 (H3N2) and A/Sydney/5/97 (H3N2)
Season	Est. excess mortality	Mean ILI per 100	$\hat{R}$ (95% CI)	Dominant virus
1998/1999	1403	1·93	1·29 (1·21–1·37)	A/Sydney/5/97 (H3N2)
1999/2000	1526	1·91	1·69 (1·07–2·45)	A/Moscow/10/99 (H3N2)
2000/2001	339	0·89	1·11 (1·07–1·15)	A/New Caledonia/20/99 (H1N1)
2001/2002	746	1·32	1·13 (1·11–1·15)	A/Panama/2007/99 (H3N2)
2002/2003	392	0·96	1·11 (0·96–1·26)	No dominant virus
2003/2004	1025	1·55	1·43 (1·08–1·83)	A/Fujian/411/2002 (H3N2)

ILI, Influenza-like illness; CI, confidence interval.
* Estimated breakpoint not significant (no outbreak).

regression-based methods; for example on data from 1976 to 1999 in the USA [8], 1990–1999 in Canada [5], and 1996–1999 in Hong Kong [9]. The US study estimated that on average the number of influenza-related deaths formed 2·2% of all deaths, while in the Canadian study the corresponding estimate was 1·9%, compared to our estimate of 2·1%. In the Hong Kong study, the estimate of influenza-related excess mortality was 16·4 deaths/100 000 population, compared to our estimate of 21·3/100 000 in the Norwegian population. These findings are not dissimilar from European studies using different methods to estimate the excess mortality: a German study for 1985–2001 [10, 11] estimated the excess mortality as 16·1 and 17·4 deaths/100 000 population for two study periods, and a Czech study for 1982–2000 [15] estimated the excess mortality as 2·2% of all deaths. Our estimates for excess mortality in the $\geqslant$65 years group of 140 and 158 deaths/100 000 are comparable with studies from Canada (108·8/100 000 population) [5], USA (132·5/100 000 population) [8], and Hong Kong (136·1/100 000 population) [9].

Most previous estimates of the reproduction number R have been made for pandemic influenza, but results from some studies are comparable to our estimates. Using a model based on the Asian 1957–1958 influenza A(H2N2) pandemic in the USA, R_0 was estimated as 1·68 [26], while other studies estimated R for inter-pandemic years as 1·39 [27], or in the range between 1·2 and 1·8 [28]. These estimates

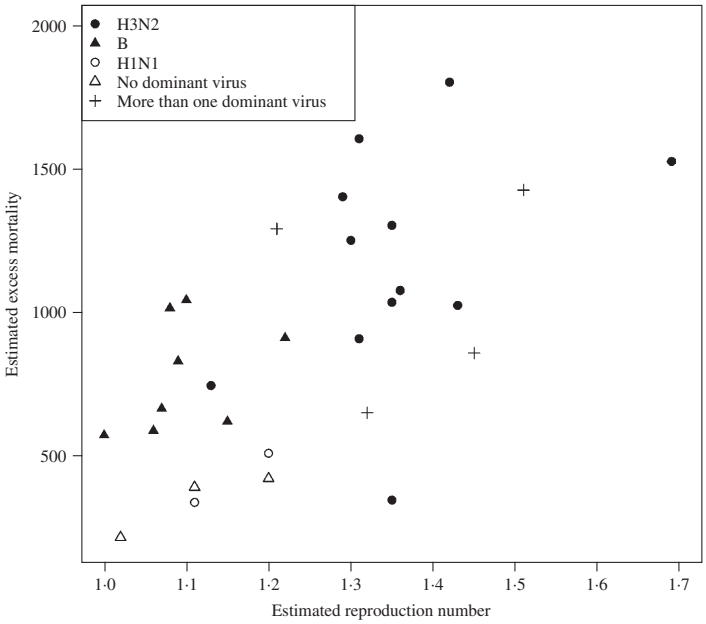

Fig. 3. Estimated excess mortality against R for each influenza season, marked by dominant virus type.

Table 2. *Regression results and estimated excess mortality for separate analyses on age groups, and for the overall analysis*

Dataset	Age group	β_{ILI}	S.E.	P value	Mean est. excess mortality	Rate per 100 000
1975–1998	0–4	0·00051	0·00019	0·0065	8	2·9
	5–14	0·00072	0·00039	0·066	3	0·6
	15–64	0·00024	0·000054	<0·0001	56	2·1
	⩾65	0·00076	0·000029	<0·0001	834	138·1
	All	0·00068	0·000027	<0·0001	911	21·7
1998–2004	0–4	0·046	0·020	0·020	8	2·6
	5–14	0·053	0·042	0·21	3	0·4
	15–64	0·0032	0·0046	0·49	13	0·4
	⩾65	0·038	0·0029	<0·0001	879	136·0
	All	0·033	0·0027	<0·0001	905	20·4

correspond well to our estimates, which ranged from about 1 in seasons with no obvious influenza outbreak to a maximum of 1·69.

The Poisson regression-based models proved suitable, adjusting for available markfor influenza activity and seasonality in week-to-week data. In classic epidemic-threshold models, the excess mortality is defined as all deaths exceeding a seasonal baseline threshold, based on years with low influenza activity [5]. This baseline can be found for instance by cyclic regression [10, 12–14]. Although seemingly robust, these models do not utilize available influenza data as well as Poisson regression-based models [11]. The differences in the Poisson regression-based models are mostly due to the type of surveillance measurements available for influenza, the types of seasonal marker, and the resolution of the data. The Norwegian data differ from the other Poisson-regression based studies mentioned by using the rate of ILI consultations from general practitioners as the marker for influenza

activity, instead of virological laboratory confirmations [8, 9] or influenza-certified deaths [5]. We found that the rate of ILI consultations, together with seasonal and year-to-year markers, modelled overall mortality satisfactorily.

In Poisson regression-based methods one can adjust for several variables in addition to different markers for influenza activity, e.g. other variables which influence mortality, and seasonal and year-to-year markers. Other possible explanatory variables include: dominant virus type, temperature, human respiratory syncytial virus (RSV) and other subcategories of overall death. Models with such additional explanatory variables and factor groups were explored.

Information about dominant virus types was better used as a supplement rather than including it in the model because viruses co-dominate and vary in magnitude between seasons. This leaves the year-to-year variability to be covered by the season variable. Temperature is often mentioned as an explanatory factor for overall death [29], but is unfeasible to use in Norway due to geographical and meteorological variability. Using the weekly average temperature in Oslo as a covariate gave a very small but significant impact; the reduction in over-dispersion was minimal and the results were broadly unchanged. For RSV infections, there are no consistent national data available, and RSV outbreaks rarely coincide with influenza outbreaks.

For our data a lag of 1 week between ILI activity and mortality gave the best model fit. A 1-week lag was also used in other similar analysis [5]. Using months instead of weeks to control for seasonality has been done in other studies [5], but in our case this weakened model fit.

Clinical surveillance data on ILI were collected from all general practices in primary care for 1975–1998 and from selected sentinels for the period 1998 onwards. Completeness of data may pose a problem for the first period for which we did not collect any denominator data. However, coverage has been estimated to be constant over the years with about 60% of the practices reporting every week. For the second period we did collect denominator data, and variability in completeness was not affected as much.

Clinically reported ILI, rather than laboratory confirmations, was chosen as a proxy for influenza activity. Laboratory diagnostics and their use have been developing over the study period and extensive data on laboratory confirmations in Norway are only available from 1999 onwards. These data show that the trends in the number of virus confirmations matched the ILI numbers well in seasons with obvious outbreaks, but in seasons with low influenza activity the fit is less good. In these seasons laboratory detections of influenza reflect not only influenza activity, but also laboratory testing activity due to outbreaks of other respiratory pathogens. As long as sampling in a population is representative, data on laboratory confirmations may serve well in some countries, but in Norway this information comes from a limited number of laboratories, and the testing practice and sensitivity varies widely from season to season and in the different laboratories. The ILI data are more robust for other factors, and the good agreement detected between ILI and mortality was not detected using laboratory data, as seen by a worse model fit when explaining mortality using laboratory data in the available period with a Poisson model. Although ILI consultation rates do not reflect the true rates of influenza in absolute numbers, they do reflect the epidemic curve. This is supported by the coincidence of peaks of ILI and excess mortality during seasons with large influenza outbreaks, as well as the agreement between peaks of ILI and the numbers of laboratory confirmations in such seasons, where data are available. The clinical definition of influenza has not changed in Norway over the years and is the same nationwide. Consequently regional and temporal biases are not expected. However, during each seasonal outbreak clinicians may be more inclined to use the R80 Influenza diagnosis when they know influenza virus is circulating, thereby enlarging the size of the peak, but the time period for the peak will not shift. Altered health-seeking behaviour over the years may affect the total number of ILI cases per season but would probably not change sufficiently quickly during a single season to alter the shape of the outbreak curve. Hence, we do not believe it would affect the relationship with mortality.

The 'epidemiological week' of the ILI surveillance was recorded from Friday to Thursday, 3 days earlier than the calendar week (in Norway: Monday–Sunday). As we tried different lag times between ILI and death, and found that 1 week gave the best fit, this should not influence the result.

The agreement of ILI and mortality in terms of peak and initial rise, and of ILI and laboratory data in seasons with obvious influenza outbreaks, suggests that ILI is also a suitable variable for estimating reproduction numbers. The Norwegian laboratory data

are not sufficiently robust for such an estimation, and, where available, will tend towards overestimation (especially for seasons with low activity). However, the *R* estimates based on ILI data have some potential sources of bias. Unexpected decrease in ILI activity was found around Christmas and New Year in some seasons, probably due to lower registration in the holidays. These artefacts would be covered by the week-to-week markers when estimating excess mortality, but could cause some underestimation of *R* in seasons where the ILI peak is around New Year. There is a potential for overestimation due to clinicians being more inclined to use the influenza diagnosis during certain periods. Estimating *R* from the initial growth of the outbreak would not be as vulnerable for such bias as methods using the entire epidemic curve in the estimation. Asynchronous influenza epidemics across Norway would also be a source of bias, but surveillance data suggests that, for most seasons, epidemics are concurrent. At worst, epidemics may be displaced by 1–2 weeks for parts of the outbreak between the most distant regions. The assumptions of an incubation and infectious period also represent some uncertainty, while the uncertainty from the regression estimating the initial growth rate is quantified through 95% CIs for *R*. However, reproduction numbers are important quantities for measuring the impact of and comparing outbreaks, and they are central in modelling the impact of counter-measures. Few estimates for seasonal influenza are available in the literature, and the estimates from the Norwegian data serve as a useful addition. Our results fit well with existing estimates, which is an interesting finding in itself, and further proves that it is possible to estimate reproduction numbers using clinically reported ILI data.

In conclusion, the Poisson regression-based methods proved useful in explaining the number of overall deaths per week by reported ILI cases. Reproduction numbers *R*, estimated using reported ILI cases, ranged up to about 1·7, supporting results found by studies using different methods. Overall, influenza was estimated to contribute to more than 2% of all deaths in Norway.

ACKNOWLEDGEMENTS

This work was supported by the Research Council of Norway, contract/grant number 170620/V30. We also thank Kristian Waalen at NIPH, Department of Virology, for compiling historical data on dominant virus strain for each influenza season, and Arve Sjølingstad at NIPH, Department of Research Data, for supplying data on all-cause mortality.

DECLARATION OF INTEREST

None.

REFERENCES

1. **Betts RF.** Influenza virus. In: Mandell GL, Bennett JE, Dolin R, eds. *Principles and Practice of Infectious Diseases*, 4th edn. New York: Churchill Livingstone, 1995, pp. 1546–1567.
2. **Heymann DL.** *Control of Communicable Diseases Manual*. Washington: American Publich Health Association, 2004.
3. **La Rosa AM, Whimbey E.** Respiratory viruses. In: Cohen J, Powderly WG, eds. *Infectious Diseases*, 2nd edn. Edinburgh: Mosby, 2004, pp. 2067–2082.
4. **Bergsaker M, Hungnes O, Iversen B.** Vaccination against influenza–why, for whom and with which vaccine? [in Norwegian]. *Tidsskrift for Norsk Laegeforening* 2006; **126**: 2814–2817.
5. **Schanzer DL, *et al.*** Influenza-attributable deaths, Canada 1990–1999. *Epidemiology and Infection* 2007; **135**: 1109–1116.
6. **Dushoff J, *et al.*** Mortality due to influenza in the United States – an annualized regression approach using multiple-cause mortality data. *American Journal of Epidemiology* 2006; **163**: 181–187.
7. **Linde A, Ekdahl K, Lindbäck J.** News on influenza and RS in Sweden [in Swedish]. *Smittskydd* 1999; **5**: 110–111.
8. **Thompson WW, *et al.*** Mortality associated with influenza and respiratory synctial virus in the United States. *Journal of the American Medical Association* 2003; **289**: 179–86.
9. **Wong C-M, *et al.*** Influenza-associated mortality in Hong Kong. *Clinical Infections Diseases* 2004; **39**: 1611–1617.
10. **Zucs P, *et al.*** Influenza associated excess mortality in Germany, 1985–2001. *Emerging Themes in Epidemiology* 2005; **2**: 6.
11. **Dushoff J.** Assessing influenza-related mortality: comment on Zucs *et al*. *Emerging Themes in Epidemiology* 2005; **2**: 7.
12. **Simonsen L, *et al.*** Impact of influenza vaccination on seasonal mortality in the US elderly population. *Archives of Internal Medicine* 2005; **165**: 265–272.
13. **Simonsen L, *et al.*** The impact of influenza epidemics on mortality: introducing a severity index. *American Journal of Public Health* 1997; **87**: 1944–1950.
14. **Simonsen L, *et al.*** Estimating deaths due to influenza and respiratory syncytial virus. *Journal of the American Medical Association* 2003; **289**: 2499–2500.
15. **Kyncl J, *et al.*** A study of excess mortality during influenza epidemics in the Czech Republic, 1982–2000. *Journal of Epidemiology* 2005; **20**: 365–371.

16. **Mamelund SE, Iversen BG.** Morbidity and mortality in pandemic influenza in Norway [in Norwegian]. *Tidsskrift for Norsk Laegeforening* 2000; **120**: 360–363.
17. **R Development Core Team.** *R: A Language and Environment for Statistical Comput-ing*. R Foundation for Statistical Computing, Vienna, Austria, 2008.
18. **Anderson RM, May RM.** *Infectious Diseases of Humans – Dynamics and Control*. Oxford: Oxford University Press, 1991.
19. **Diekmann O, Heesterbeek JA.** *Mathematical Epidemiology of Infectious Diseases. Model Building, Analysis and Interpretation*. Chichester: Wiley, 2000.
20. **Chowell G, Nishiura H, Bettencourt LMA.** Comparative estimation of the repro-duction number for pandemic influenza from daily case notification data. *Journal of the Royal Society Interface* 2007; **4**: 155–66.
21. **Gran JM, *et al.*** Growth rates in epidemic models: application to a model for HIV/AIDS progression. *Statistics in Medicine* 2008; **27**: 4817–34.
22. **Favier, C, *et al.*** Early determination of the reproductive number for vector-borne diseases: the case of dengue in Brazil. *Tropical Medicine & International Health* 2006; **11**: 332–40.
23. **Zeileis A, *et al.*** strucchange: an R package for testing for structural change in linear regression models. *Journal of Statistical Software* 2002; **7**: 1–38.
24. **Chowell G, *et al.*** The basic reproductive number of Ebola and the effects of public health measures: the cases of Congo and Uganda, *Journal of Theoretical Biology* 2004; **229**: 119–126.
25. **Wearing HJ, Rohani P, Keeling MJ.** Appropriate models for the management of infectious diseases. *PLoS Medicine* 2005; **2**: 621–627.
26. **Longini IM, *et al.*** Containing pandemic influenza with antiviral agents. *American Journal of Epidemiology* 2004; **159**: 623–633.
27. **Gani R, *et al.*** Potential impact of antiviral drug use during influenza pandemic. *Emerging Infectious Diseases* 2005; **11**: 1355–1362.
28. **Viboud C, *et al.*** Transmissibility and mortality impact of epidemic and pandemic influenza, with emphasis on the unusually deadly 1951 epidemic. *Vaccine* 2006; **24**: 6701–6707.
29. **Laake K, Sverre JM.** Winter excess mortality: a comparison between Norway and England plus Wales. *Age and Aging* 1996; **25**: 343–348.

3

Analysing direct and indirect effects of treatment using dynamic path analysis applied to data from the Swiss HIV Cohort Study

Kjetil Røysland[1, *] Jon Michael Gran[1] Bruno Ledergerber[2]
Viktor von Wyl[2] James Young[3] Torben Martinussen[4]
Odd O. Aalen[1]

[1] *Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, Norway.*
[2] *Division of Infectious Diseases and Hospital Epidemiology, University of Zurich, Switzerland.*
[3] *Basel Institute for Clinical Epidemiology and Biostatistics, University Hospital Basel, Switzerland.*
[4] *Department of Biostatistics, Faculty of Health Sciences, University of Southern Denmark, Denmark.*
[*] *Correspondence to: Kjetil Røysland, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo, P.O.Box 1122 Blindern, 0317, NORWAY. Telephone: (+47)22851015, fax: (+47)22851313, e-mail: kjetil.roysland@medisin.uio.no.*

Abstract

When applying survival analysis, such as Cox regression, to data from major clinical trials or other studies, often only baseline covariates are used. This is typically the case even if updated covariates are available throughout the observation period, which leaves large amounts of information unused. The main reason for this is that such time-dependent covariates often are internal to the disease process, as they are influenced by treatment, and therefore lead to confounded estimates of the treatment effect. There are, however, methods to exploit such covariate information in a useful way. In this paper we study the method of dynamic path analysis applied to data from the Swiss HIV Cohort Study. To adjust for time-dependent confounding between treatment and the outcome 'AIDS or death', the analysis is done on a sequence of mimicked randomized trials constructed from the original cohort data. To analyse these trials together, regular dynamic path

analysis is extended to a composite analysis of weighted dynamic path models. Results using a simple path model, with one indirect effect mediated through current RNA level, show that most, or all, of the total effect go through RNA for the first four years after treatment start. A similar model, but with CD4 level as the mediating variable, show a weaker indirect effect, but the results are i the same direction. There are many reasons to be cautious about making to conclusive statements about estimates of direct and indirect effects, such as the ones from dynamic path analyses. Dynamic path analysis is however a useful tool to explore underlying processes which are ignored in regular analyses. Such analyses of direct and indirect effects also usually have a more provisional status, and may not require the same level of certainty.

1 Introduction

Survival analysis has established itself as a major tool in clinical research. Recently, it was demonstrated to be the most commonly used of the more advanced statistical methods in the New England Journal of Medicine [1]. However, a surprising aspect of survival analysis as practiced in medical applications and other fields is the lack of attention to covariate information collected between the starting point and the event. Commonly, one sees major clinical trials and other studies where the main analysis is a Cox regression using only baseline covariates; this in spite of the fact that many of these covariates are updated repeatedly throughout the period of observation. Clearly, large amounts of useful information are not being used.

It is not hard to understand the reasons behind this situation. Typically, these covariates will be internal to the disease process; they will be influenced by a possible treatment effect, and therefore they will, so to speak, 'steal' treatment effect from the event analysis. Thus, when a time-dependent Cox model is run the treatment effect will typically be underestimated. This is a very well known problem and one is thoroughly warned against it in the survival analysis literature, see e.g. [2]. Nevertheless, it is a waste of information to ignore time-dependent covariates and there exist procedures for exploiting this information, see e.g. dynamic

path analysis as described in [3, 4].

More fundamentally, the skeptical attitude to detailed analysis of time-dependent covariates is that this pertains to the issue of how the treatment effect is mediated towards its final goal of (possibly) prolonging survival. More, precisely, in a dynamic path analysis we attempt to estimate indirect effects of treatment mediated through one or more markers, or time-dependent covariates. Such mediation problems are vulnerable to unmeasured confounders, and are in general dependent on some prior mechanistic understanding. Nevertheless, such analysis could contribute in a valuable way to understanding the mechanisms behind treatment effects.

It is important to realize that in biomedical research, mechanistic understanding is typically limited. A new medication, for example, may be suggested because the mechanistic knowledge suggests that it should have a positive effect. However, one can never be sure that it lives up to expectations. There are numerous examples of treatments that have reached even the level of phase III clinical trials, but do not yield the expected results. See, for example, Rossebø *et al.* [5] where extensive cholesterol reduction does not achieve the aim of limiting the progression of aortic stenosis. One reason for such a negative result is that biological pathways are very complex. Medical interventions, such as taking a medication, may typically have the effect of blocking certain receptors, that is, of blocking certain paths in the pathway system. However, the effect of such interventions will tend to have many types of effects, following many pathways and the net result may not be as expected.

The uncertainty within mechanistic understanding of biological systems means that one should use all statistical information available to throw more light on mechanistic issues. This includes, for example, using repeated measurements taken during the study in clinical survival studies. This may contribute to understanding, even though the causal issues may not be clarified.

In the present paper we consider a situation more complex than standard survival analysis. Analysing HIV cohort data, we are faced with the fact that treatment start is dependent on the state of the patient; treatment is rarely started unless the infection has shown progression by effecting the immune system, measured for example by the CD4 cell count. This creates time-dependent confound-

ing between treatment and CD4 count. We use a method for resolving this issue, based on mimicking a series of randomized trials from the original cohort data, proposed in Gran *et al.* [6]. We then study how dynamic path analysis can be used together with these ideas, to gain more insight into how the treatment effect is mediated.

Dynamic path analysis is an extension of classical path analysis, where the variables can be time-dependent. The final outcome is a stochastic process, typically a counting process in a survival or event history setting. Dynamic path analysis can be used to describe how the effect of a fixed covariate, such as treatment, works directly and partly indirectly, through time-dependent covariates. The aim is to decompose the total effect of treatment into a sum of indirect effects, which are mediated through time-dependent covariates, as well as the direct effect.

Note, as stated in [7], that there exists no unique definition of the concepts of direct, indirect and total effects, and that unmeasured confounders could cause problems to the interpretation of such concepts. Nevertheless, dynamic path analysis is a useful tool for understanding how the treatment effect on a time to event endpoint is mediated through time-dependent covariates.

When analysing many mimicked randomized trials, composite likelihood inference [8–10] serve as a simple and efficient method for combining the different analyses, and giving overall effect estimates. The parameters based on such inference would also keep their interpretation, given certain assumptions.

In Section 2 we describe the dataset from the Swiss HIV Cohort study, while the approach for constructing mimicked randomized trials is described in Section 3. The dynamic path model and path effects are presented in Section 4, while the estimators for these path effects follow in Section 5. In section 6 we apply our methods to the Swiss cohort data, with a discussion following in Section 7.

2 The Swiss HIV cohort dataset

The methods in this paper are applied to data from the Swiss HIV Cohort Study [13], which is an ongoing multi-center research project following up HIV infected adults aged 16 or older. The data begins in January 1996, when Highly active antiretroviral treatment (HAART) became available in Switzerland, and

run until September 2003. The time scale used is months since the start of follow up, with baseline being the time of the first visit after January 1996. Patients who died or refused further participation before 1996, who were on HAART or in in clinical stage C at baseline, or whose treatment history before joining the cohort was uncertain were excluded. Time between visits varies, but scheduled clinical follow-up is every sixth month, with additional laboratory measures taken every third month. On months without visits, the last observation is carried forward. The variables measured include CD4 count, HIV-1 RNA and haemoglobin levels, together with indicator variables describing whether the individuals have been treated with mono therapy, dual therapy or HAART, or experienced a CDC stage B event (a disease associated with HIV but less severe than an AIDS defining disease). When individuals first start treatment they are considered as on treatment from that time and until the end of the study.

The dataset consists of data from 2161 individuals, who were observed over a maximum of 92 months. The total number of person-months of observation is 77 838. Two hundred and two of the individuals progressed to AIDS or death, while 717 were never treated with HAART. The dataset used is the same as was analyzed in Sterne *et al.* [14] and in Gran *et al.* [6].

3 Mimicking randomized trials

3.1 Constructing a proper pseudo dataset

When analysing the Swiss HIV Cohort data we handle the problem of time-dependent confounding, represented by variables such as the HIV-1 RNA and CD4 cell count, by analysing a sequence of mimicked randomized trials, as was done in Gran *et al.* [6].

Each mimicked trial is constructed based on individuals starting treatment in a certain time interval or, as in our situation, in a certain month since start of follow-up. Individuals starting treatment in that particular month serve as the treatment group, while the individuals not yet on treatment serve as the control group. To avoid confounding due to treatment, individuals in the control group are artificially censored at the time of later treatment start. Only baseline

covariates and covariates at the start of each mimicked trial are controlled for in the analysis of each such subset of the real data. Dependent censoring, which might be introduced when constructing the pseudo datasets, as well as other dependent censoring, is adjusted for using inverse probability of censoring (IPC) weights.

The purpose of the weighting procedure is to produce a weighted dataset which reflects most of the mechanisms of the unweighted dataset, but for which the censoring can now be considered as independent.

Let A be an indicator for initiating treatment at the start of the mimicked trial (valued 1 if treated and 0 otherwise), $C(t)$ an indicator for the combined event 'censoring or artificial censoring' at time t (1 if censored and 0 otherwise), $\boldsymbol{Z}$ a vector of time-independent baseline covariates, $\boldsymbol{L}(t)$ a vector of time-dependent covariates at time t, and $\bar{\boldsymbol{L}}(t)$ the time-dependent covariate history up to time t. The IPC weight $w_i(t)$ for individual i at time t, weighting for both regular censoring and the artificial censoring done when creating the pseudo dataset, can be constructed as the IPC weights in Robins *et al.* [11] and Hernan *et al.* [12], using

$$w_i(t) = \sum_{k=0}^{t} \frac{P(C(k) = 0 | \bar{C}(k-1) = 0, A = a_i, \boldsymbol{Z} = \boldsymbol{z}_i)}{P(C(k) = 0 | \bar{C}(k-1) = 0, A = a_i, \bar{\boldsymbol{L}}(k-1) = \bar{\boldsymbol{l}}_i(k-1))}, \qquad (1)$$

where by definition $\bar{C}_a(-1) = 0$.

Let $\mu_i(s)$ be the censoring intensity for individual i, conditioned on its full covariate history. If the covariates and events are observed in discrete time intervals, where t_k is the end time of the k^{th} interval, we can approximate the denominator in (1) by

$$P(C(k) = 0 | \bar{C}(k-1) = 0, A = a_i, \bar{\boldsymbol{L}}(k-1) = \bar{\boldsymbol{l}}_i(k-1)) \approx \exp(-\int_{t_{k-1}}^{t_k} \mu_i(s) ds).$$

Correspondingly, given that $\tilde{\mu}_i(s)$ is the censoring intensity for individual i conditioning on its *marginal* covariate history, we can approximate the numerator in

(1) by

$$P(C_a(k) = 0 | \bar{C}_a(k-1) = 0, A = a_i, \boldsymbol{Z} = \boldsymbol{z}_i) \approx \exp(-\int_{t_{k-1}}^{t_k} \tilde{\mu}_i(s) ds).$$

In other words, we can use that

$$w_i(t) \approx \exp(-\int_o^t \mu_i(s) - \tilde{\mu}_i(s) ds),$$

where $\mu_i(s)$ and $\tilde{\mu}_i(s)$ can be estimated using Aalen's additive regression model [4].

3.2 Interpretation of the censoring weights

Imagine that the unweighted data for a mimicked randomized trial is distributed according to some probability measure P. As mentioned in Section 3.1, the censoring done to create this dataset could be dependent, leading to biased estimates. We adjust for such bias using the weights $w_i(t)$. It is natural to interpret these weights as a likelihood ratio process.

More precisely, we consider a new probability measure Q that is absolutely continuous with respect to P. Q differs from P in that the censoring is independent, but the behavior of the remaining processes is left unchanged. The likelihood ratio between the measure P and Q, over the history before time t, gives the weight at time t. It is interesting to note that our approximation of the weights suggested by Robins *et al.* [11] coincide asymptotically with a special case of the well known Jacod formula for likelihood ratios. These aspects of data re-weighting are discussed in [15].

4 Dynamic path model and path effects

If the censoring is independent, then the mimicked randomized trials in Section 3.1 can be analysed using dynamic path analysis as it was proposed by Fosen *et al.* [3]. Let us start with the analysis of one such mimicked trial.

We will consider a simple dynamic path model, a model with one direct and one indirect path going from treatment to the main event 'AIDS or death'. The

indirect path is mediated through a time-dependent continuous covariate. We will in two separate analyses consider models with both the RNA and CD4 level as the mediating variable. The RNA variable gives a measure of the amount of HIV-1 RNA in the blood of the HIV infected individual, while the CD4 variable measure the number of CD4 immune cells.

Let A_i be the treatment indicator for individual i, $L_i(t)$ the level of the mediating variable (either the RNA or CD4 level) for individual i at time t, $\boldsymbol{Z}_i$ a vector of relevant baseline covariates for individual i, and $D_i(t)$ an indicator function, 1 if individual i dies or is diagnosed with AIDS at time t, and 0 otherwise. We have n independent observations A_i, $L_i(t)$, $\boldsymbol{Z}_i$, and $D_i(t)$, of A, $L(t)$, $\boldsymbol{Z} = (Z_1, ..., Z_p)$, and $D(t)$, where p is the number of baseline covariates. The model is illustrated by the dynamic path diagram [3] in Figure 1.

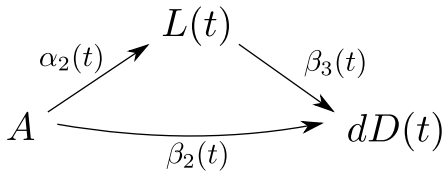

Figure 1: Path diagram for the dynamic model with a direct path from treatment A to the increment of the event process $dD(t)$ (AIDS or death) and an indirect path from A to $dD(t)$ through a mediating variable $L(t)$, being either the HIV-1 RNA level or the CD4 cell count. $\alpha_2(t)$, $\beta_2(t)$ and $\beta_3(t)$ are regression coefficients.

Considering the path diagram, we observe two paths going from treatment A to the event increment $dD(t)$; one direct path, and one indirect path through the time-dependent covariate $L(t)$. As in [3], the aim is to analyse the effect of the covariate processes on the infinitesimal changes of the event process $D(t)$, $dD(t)$. The dynamic path analysis is carried out using recursive least squares regression, as in regular path analysis. Each node in the diagram is regressed onto its parent, for each time t when data is potentially collected. As previously mentioned, if data for a time-dependent covariate is not available, the last observation is carried forward, as in [14]. In contrast to regular path analysis, the regression of the event $D(t)$ onto its parents is done using the additive regression model. Direct and

indirect effects can now be calculated, preserving the additivity of classical path analysis, allowing total effects to be decomposed into direct and indirect effects. An indirect effect is simply defined as the product of the regression functions (ordinary linear or additive) along its path.

The model in Figure 1 is formally given by

$$L(t) = \alpha_1(t) + \alpha_2(t)A + \alpha_3(t)\boldsymbol{Z} + M_1(t) \tag{2}$$

for all $t \in \{1, 2, ..., 92\}$, and

$$D(t) = \int_0^t Y(s)\Big(\beta_1(s) + \beta_2(s)A + \beta_3(s)L(s-) + \beta_4(s)\boldsymbol{Z}\Big)ds + M_2(t), \tag{3}$$

where $Y(t)$ is the at risk indicator at time t (valued 1 if at risk and 0 otherwise), $M_h(t)$ are martingales corresponding to the residuals, $\alpha_j(t)$ for $j \in \{1, 2, 3\}$ are regression coefficients at time t, and $\beta_k(t)$ for $k \in \{1, 2, 3, 4\}$ are regression functions. See for example Diggle *et al.* [17] for other models for longitudinal data formulated using martingale residuals.

The parameters $\boldsymbol{\alpha}(t) = (\alpha_1(t), \alpha_2(t), \alpha_3(t))$ in model (2) can be estimated using ordinary linear regression at each time $t \in \{1, 2, ..., 92\}$, and the parameters $\boldsymbol{\beta}(t) = (\beta_1(t), \beta_2(t), \beta_3(t), \beta_4(t))$ in model (3) using additive hazard regression. We define the two cumulative path effects

$$A \to D : \int_0^t \beta_2(s)ds \tag{4}$$

and

$$A \to L \to D : \int_0^t \alpha_2(s)\beta_3(s)ds \tag{5}$$

as the direct effect of treatment on the outcome, $(A \to D)$, and the indirect effect of treatment going through the variable L, $(A \to L \to D)$. In other words, direct and indirect effects along paths such as the ones in Figure 1 are found by multiplying the regression coefficients along that path. In Fosen *et al.* [3], it is shown that these path effects add up to the marginal (only baseline adjusted) effect.

5 Estimators for the path effects

5.1 Weighted dynamic path analysis

If it was not for the censoring, we could proceed as in [3] to introduce estimators for the path effects analysing one mimicked randomized trial alone. We will, however, do a weighted analysis to adjust for any dependent censoring, which corresponds to using data distributed according to P to do inference for Q (remembering the notation of Section 3.2).

Let us first introduce some more notation. Let $\boldsymbol{N}_k(t)$ be a vector where the i^{th} element $N_{k,i}(t)$ is the number of events (indicated by $D_i(t) = 1$) up until time t for individual i in trial k. $\boldsymbol{X}_k(t)$ is a matrix where the i^{th} element, $\boldsymbol{X}_{k,i}(t) = (1, A_{k,i}, \boldsymbol{L}_{k,i}^{\mathrm{T}}(t-), \boldsymbol{Z}_{k,i}^{\mathrm{T}})$, is a vector with all covariate values at time t for individual i in trial k and $\boldsymbol{Y}_k(t)$ is a vector where the i^{th} element, $Y_{k,i}(t)$, is the 'at-risk' indicator at time t for individual i in trial k.

Consider the additive hazard regression model from Equation 3 with respect to Q. We have a left continuous function $\boldsymbol{B}(t) = (\int_o^t \beta_1(s)ds, \int_0^t \beta_2(s), \beta_3(s), \beta_4(s))$, such that

$$\boldsymbol{M}_k(t) = \boldsymbol{N}_k(t) - \int_0^t \boldsymbol{Y}_k(s)\boldsymbol{X}_k(s)d\boldsymbol{B}(s)$$

is a Q-martingale with respect to the history of the individuals in trial k.

Let $\boldsymbol{W}_k(t)$ be a diagonal matrix where the i^{th} element along the diagonal $W_{k,i}(t)$ is the censoring weight at time t for individual i in trial k, and recall the discussion in 3.2. Since $\boldsymbol{W}_k(t)$ is the likelihood ratio process, we obtain from the integration-by-parts formula for stochastic integrals that

$$\int_0^t \boldsymbol{W}_k(s-)d\boldsymbol{N}_k(s) - \int_0^t \boldsymbol{W}_k(s-)\boldsymbol{Y}_k(s)\boldsymbol{X}_k(s)d\boldsymbol{B}(s)$$

defines a P-martingale. This suggests that some reasonable estimator for $\boldsymbol{B}(t)$, say $\widehat{\boldsymbol{B}}(t)$, is given by each of the estimating equations

$$\boldsymbol{X}_k(t)^T\boldsymbol{W}_k(t-)d\boldsymbol{N}_k(t) - \boldsymbol{X}_k(t)^T\boldsymbol{W}_k(t-)\boldsymbol{Y}_k(t)\boldsymbol{X}_k(t)d\widehat{\boldsymbol{B}}(t) = 0 \tag{6}$$

for every t.

For each trial k and every jump time t of $\boldsymbol{L}_k$, and for the mediating covariate

processes $\boldsymbol{L}_k(t)$, we have the linear model

$$\boldsymbol{L}_k(t) = \boldsymbol{X}_k(t)\boldsymbol{\alpha}(t) + \boldsymbol{\epsilon}_k(t),$$

where $\boldsymbol{\epsilon}_k(t)$ has zero mean and its components are uncorrelated with respect to Q.

Now,

$$\boldsymbol{W}_k(t)\Big(\boldsymbol{L}_k(t) - \boldsymbol{X}_k(t)\boldsymbol{\beta}(t) + \boldsymbol{\epsilon}_k(t)\Big)$$

has zero mean and its components are uncorrelated with respect to Q.

The corresponding weighted normal equation is

$$\boldsymbol{X}_k(t)^T\boldsymbol{W}_k(t)\boldsymbol{L}_k(t) - \boldsymbol{X}_k(t)^T\boldsymbol{W}_k(t)\boldsymbol{X}_k(t)\widehat{\boldsymbol{\beta}}(t) = 0. \tag{7}$$

Estimates of $\boldsymbol{\alpha}(t)$ and $\boldsymbol{B}(t)$, $\hat{\boldsymbol{\alpha}}(t)$ and $\hat{\boldsymbol{B}}(t)$, inserted into equation (4) and (5) suggests the estimators

$$A \to D : \widehat{B}_2(t) \tag{8}$$

and

$$A \to L \to D : \int_0^t \widehat{\alpha}_2(s)d\widehat{B}_3(s) \tag{9}$$

for the cumulative path effects at time t. $\widehat{B}_l(t)$ is the l^{th} entry of $\widehat{\boldsymbol{B}}(t)$.

5.2 Estimating composite treatment effects combining all mimicked trials

Until now we have considered the dynamic path analysis of one mimicked randomized trial, where in the last section we extended the dynamic path analysis from [3] by introducing censor weights. However, aiming to analyse the full cohort dataset, we construct many mimicked randomized trials; one for each possible time of treatment start. We therefore seek to combine many such weighted dynamic path analyses, in a similar way as was done with the Cox regressions in [6], by using composite likelihood inference.

Composite additive hazard regressions

Let us first combine the additive regressions done for each mimicked trial. In order to combine all the mimicked trials to get an overall estimate, we aggregate over the estimating equations (6), resulting in the the estimating equation

$$\sum_k \boldsymbol{X}_k(s)^T \boldsymbol{W}_k(s-) d\boldsymbol{N}_k(s) - \sum_k \boldsymbol{X}_k(s)^T \boldsymbol{W}_k(s-) \boldsymbol{X}_k(s) d\widehat{\boldsymbol{B}}(s) = 0.$$

A short matrix computation shows that the corresponding aggregated estimator is

$$\widehat{\boldsymbol{B}}(t) = \sum_k \int_0^t \left(\sum_l \boldsymbol{X}_l(s)^T \boldsymbol{W}_l(s-) \boldsymbol{X}_l(s) \right)^{-1} \boldsymbol{X}_k(s)^T \boldsymbol{W}_k(s-) d\boldsymbol{N}_k(s). \tag{10}$$

Composite linear regressions

Correspondingly, we must aggregate the weighted normal equations (7) for all observed time points in each mimicked trial. This gives the aggregated estimation equation

$$\sum_k \boldsymbol{X}(t)_k^T \boldsymbol{W}_k(t) \boldsymbol{L}(t)_k - \boldsymbol{X}(t)_k^T \boldsymbol{W}_k(t) \boldsymbol{X}(t)_k \widehat{\boldsymbol{\beta}}(t) = 0,$$

and the estimator

$$\widehat{\boldsymbol{\beta}}(t-) = \sum_k \left(\sum_l \boldsymbol{X}(t)_l^T \boldsymbol{W}_l(t-) \boldsymbol{X}(t)_l \right)^{-1} \boldsymbol{X}(t)_k^T \boldsymbol{W}_k(t-) \boldsymbol{L}(t)_k. \tag{11}$$

Estimating cumulative path effects

The estimators (10) and (11) can now be used in (8) and (9), giving estimators for the aggregated path effects. Confidence intervals, or more precisely percentile intervals, are calculated using bootstrap re-sampling methods [18], re-sampling at an individual level on all the data used in the composite analysis.

6 Results

6.1 Mimicked randomized trials constructed from the Swiss HIV Cohort data

Using the data from the Swiss HIV Cohort Study we construct 92 mimicked randomized controlled trials, as in [6]. Figure 2 show the mean levels of RNA and CD4, together with individual trajectories, for the treatment and control group in the first mimicked trial. The choice of trial is arbitrary, and other trials generally show a similar trend.

Figure 2 illustrates how the RNA levels quickly decrease for individuals receiving treatment, before stabilizing. The CD4 level for individuals on treatment increases, but does not stabilize as fast as the RNA level. In the control group, the RNA and CD4 levels are quite stable. Note that the RNA and CD4 variables used in the following analyses take the same form as in Figure 2. The RNA level is measured as the logarithm of HIV-1 RNA copies per millilitre and the CD4 level is the square root of CD4 cells per microlitre.

6.2 Results using a model with an indirect treatment effect through the RNA level

Let us first consider the model in Section 4 with RNA level at time t as the mediating variable, and perform a composite weighted dynamic path analysis of all the mimicked randomized trials. We now seek to estimate the indirect effect of treatment through RNA, together with the remaining direct effect, using the estimators from Section 5.2. As noted earlier, the indirect and direct effect add up to the total effect. The baseline covariates used, corresponding to the vector $\boldsymbol{Z}$ in the equations, include sex, risk group, age at baseline and the following covariates at inclusion, at the start of the mimicked trial and lagged values at the start of the mimicked trial: CD4 group (grouped into 0-49, 50-99, 100-199, 200-349, 350-499, 500-749, $\geq$ 750 cells per μL), RNA group (grouped into <400, 400-1000, 1001-10 000, 10 001-100 000, >000 000 copies per mL), haemoglobin group (grouped into fifths), CDC B event and previously experienced CDC B event.

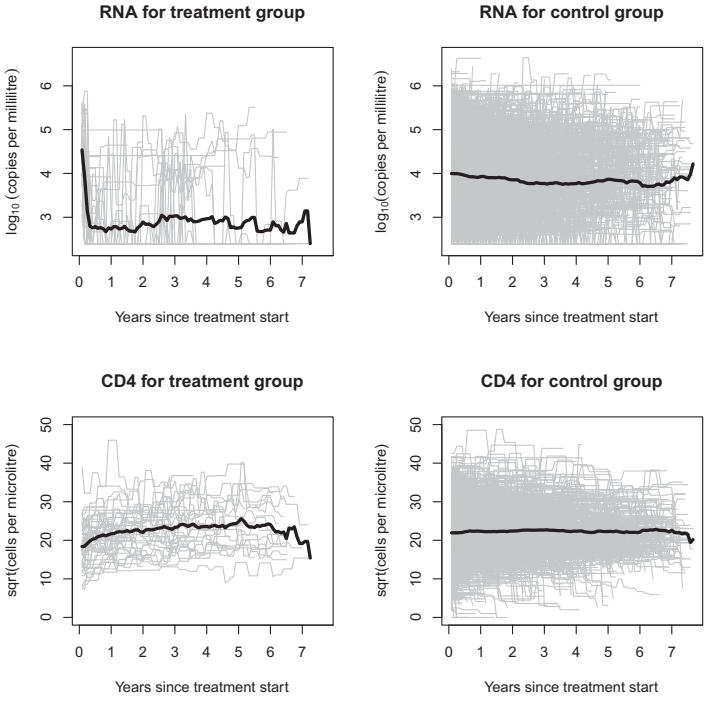

Figure 2: RNA and CD4 levels for individuals starting treatment and the controls in the first mimicked randomized controlled trial. The black line is the mean RNA or CD4 level at a certain time point, while gray lines indicate individual trajectories.

Lagged values are values three months before, corresponding to the scheduled time between visits [14]. All programming was done in the statistical package R, version 2.10.1 [19]. Estimated cumulative effects, and the corresponding regression functions (not cumulative), are plotted in Figure 3.

The first three panels of the figure, with plots of the total effect and the two path effects, show that there is a fairly constant indirect effect of treatment going through RNA for the first 4 years after treatment start. After that, the effect decreases. The direct effect of treatment is close to zero for the first 3-4 years,

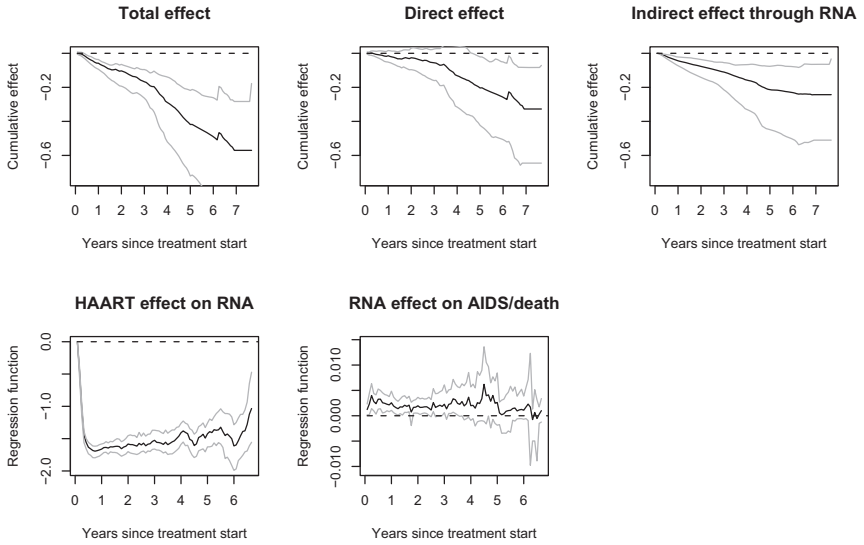

Figure 3: Estimated cumulative direct and indirect effects, and regression functions, from the composite weighted dynamic path analysis using the model in Figure 1. Grey lines represent 95% percentile intervals based on 100 bootstrap replications. Note that the last two panels are regression functions, which are not cumulative, corresponding to the parameters $\alpha_2(t)$ and $\beta_3(t)$ respectively.

before there is a corresponding shift towards an increasing direct effect. We see that the uncertainty, represented by the 95% percentile intervals, grows larger to the right as fewer people are left observed.

In the two last panels of Figure 3 we see the regression functions for the treatment effect on RNA and the RNA effect on the event 'AIDS or death', or $\alpha_2(t)$ and $\beta_3(t)$ from the equations in Section 4 and in Figure 1. We see that the effect of treatment on RNA appears to be immediate, and then it remains constant, or possibly slowly decreases. The RNA effect on the event 'AIDS or death' on the other hand, is fairly constant. The regression functions in the two last panels are cut after 80 months, to remove the noise from the uncertainty at the end to better see the development in the first 80 months.

6.3 Results using a model with an indirect treatment effect through the CD4 level

Let us then consider the same model as in the previous section, but with CD4 level as the mediating variable instead of RNA. The results from a composite weighted dynamic path analysis is presented in Figure 4. As in the previous figure, the three top panels show the cumulative total, direct and indirect effect, while the bottom two panels show the regression functions $\alpha_2(t)$ and $\beta_3(t)$ from the equations in Section 4.

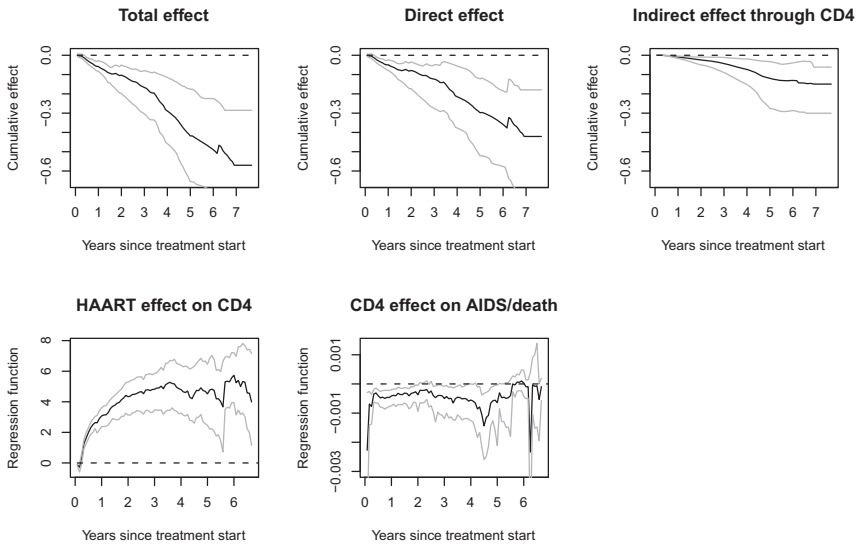

Figure 4: Estimated cumulative direct and indirect effects, and regression functions, from the composite weighted dynamic path analysis using the model in Figure 1. Grey lines represent 95% percentile intervals based on 100 bootstrap replications. Note that the last two panels are regression functions, which are not cumulative, corresponding to the parameters $\alpha_2(t)$ and $\beta_3(t)$ respectively.

At first glance, the results from Figure 4 are very similar to the results using RNA as the intermediate variable. We see, however, that the indirect effect of

treatment going through the CD4 level is smaller than the indirect effect through RNA shown in Section 6.2. We also see that the regression functions for HAART on CD4 and CD4 on the final outcome 'AIDS or death' behave differently than for the corresponding regression functions using RNA in the previous section. An obvious difference between RNA and CD4 is that CD4 levels will decrease while RNA levels will increase with worsened disease progression, but there is also differences in how these levels behave. The treatment effect on CD4 from Figure 4 seems to be working slower than the effect on the RNA level and stabilizing much later. The regression functions in the two last panels are again cut after 80 months.

7 Discussion

From the results analysing the Swiss HIV Cohort data in Section 6, we see that using HIV-1 RNA level as the mediating variable in our dynamic path model captures more of the total effect than using the CD4 cell count. The direct effect plotted in Figure 3 actually suggests that most, or all, of the treatment effect can be explained by the RNA level for the first three or four years after starting treatment. After that, the indirect effect of RNA levels off, and the direct effect increases. One interpretation of this would be that the treatment effect is explained by the RNA level for the first three or four years after treatment start, while after that other factors become more important. An explanation to why the effect decreases after three or four years could be that patients on treatment are more closely clinically monitored, and also more likely to receive prophylaxis against opportunistic infections if indicated. Also, patients not (yet) on treatment may have special characteristics with poorer outcomes (e.g. injecting drug use and presumed poor treatment adherence). However, HAART, and especially protease inhibitors, appear to have a number of effects that are not necessarily mediated through the effect on viral replication [20–22], so it is not obvious that all effects are mediated through viral load.

The estimated direct and indirect effect using CD4 level as the mediating variable, found in Figure 4, point in the same direction as for the estimates using RNA, but the effect going through CD4 is smaller. From the plots of the regression

functions in Figure 3 and Figure 4 of HAART effect on RNA and CD4, we see that treatment affects the RNA level much more rapidly than the CD4 level. Both levels stabilize a certain time after treatment start, but RNA stabilizes much faster. This could be part of the explanation as to why the indirect effect through RNA is greater than the one through CD4. The same propensity to faster change in RNA levels is also seen in the descriptive plots in Figure 2.

Note that when interpreting plots like the ones in Figure 3 and Figure 4, i.e., both time-dependent effect estimates and estimated regression functions, the uncertainty will increase towards the right because fewer observations are available. This is reflected in wider confidence intervals. In other words, the most reliable information is found to the left in these figures.

There are reasons to be cautious about making conclusive statements about estimates of direct and indirect effects, such as those estimated using dynamic path analysis. The most important reason could be that any possible confounding between the mediator and the final outcome will affect the estimates. When estimating total effects, one needs to assume no unmeasured confounding between exposure and the outcome; and when estimating direct and indirect effects, there should also be no unmeasured confounding between mediator and outcome [23]. These assumptions will therefore be harder to satisfy, and the problem will increase with the number of mediators. As when estimating total effects, assumptions of no unmeasured confounders are generally not testable when estimating direct and indirect effects. See [24], where additional assumptions for effect decomposition, beyond the absence of confounders, are discussed. Note also that measurement error in mediating variables often will cause indirect effects to be underestimated [25]. Another important point is that the estimated indirect and direct effects are based on a specified model, modelling how the variables affect each other, and will therefore depend on whether this model is a correct description of the actual system of variables. This will also include how the variables are represented, for example if they should be parameterized or dichotomized to be appropriate in a linear model. As for the method of dynamic path analysis, one should note that this still is a rather novel approach and that there remains a need to formalize certain aspects of its use [7].

The main purpose of this paper is to explore methodological possibilities of

dynamic path analysis. The present study is based on a rather limited number of endpoint from the early years of antiretroviral treatment, suggesting that it is premature to draw solid biological and clinical conclusions. However, despite the reservations, there are still much insight to gain from dynamic path analyses of direct and indirect effects, including insight on processes which are ignored in analyses of total effects. One could argue that studies about direct and indirect effects usually have a more provisional status, and should not require the same level of certainty as for studies of total effects.

The dynamic path models used to analyse the data in this paper are simple models including only one intermediate variable included at a time, namely the level of RNA copies or CD4 cells. One would expect that antiretroviral treatment would lower the RNA level, and have a positive effect on the outcome 'AIDS or death'. Similarly, one would expect to see similar effect if the RNA level was replaced with CD4 level, which is the most commonly used marker for HIV disease progression. Biologically, one would expect that the effect of antiretroviral treatment on clinical progression is channeled via a path going through the RNA level and then CD4 level. This would suggest more complex models, with both RNA and CD4 included at the same time and with several possible indirect paths going from treatment to the outcome. However, models with such paths would demand more from the data to capture necessary details. Effect estimates for such complex paths will, for example, be more easily subject to confounding and harder to interpret. The fact that laboratory measures are taken every third month at best may weaken attempts to model the interaction between variables, such as RNA and CD4 levels. We therefore chose to focus on two simple dynamic path models in this paper, to demonstrate how the idea of mimicking randomized trials and doing composite analyses in [6] can be used to extend the method of dynamic path analysis and explore datasets such as the Swiss HIV Cohort data. We have shown that even the most simple dynamic path models can be used to gain more insight about underlying processes in such datasets.

Funding

This work was supported by the Research Council of Norway, contract/grant number: 170620/V30.

References

1. Horton NJ, Switzer SS. Statistical methods in the Journal. *New England Journal of Medicine* 2005; **353**:1977–1979.

2. Prentice RL, Kalbfleisch JD. Mixed discrete and continuous Cox regression model. *Lifetime Data Anal.* 2003; **9**(2):195–210. DOI: 10.1023/A:1022935019768

3. Fosen J, Ferkingstad E, Borgan Ø, Aalen OO. Dynamic path analysis – a new approach to analyzing time-dependent covariates. *Lifetime Data Anal* 2006; **12**:143–167. DOI: 10.1007/s10985-006-9004-2

4. Aalen OO, Borgan Ø, Gjessing HK. *Event History Analysis, A Process Point of View*. 2008.

5. Rossebø AB, Pedersen TR, Boman K, et al. Intensive lipid lowering with simvastatin and ezetimibe in aortic stenosis. *New England Journal of Medicine* 2008; **359**:1343–1356. DOI: 10.1056/NEJMoa0804602

6. Gran JM, Røysland K, Wolbers M, Didelez V, Sterne JAC, Ledergerber B, Furrer H, von Wyl V, Aalen OO. A sequential Cox approach for estimating the causal effect of treatment in the presence of time-dependent confounding applied to data from the Swiss HIV Cohort Study. *Statistics in Medicine* 2010. DOI: 10.1002/sim.4048

7. Martinussen T. Dynamic path analysis for event time data: large sample properties and inference. *Lifetime Data Anal* 2010; **16**:85–101. DOI: 10.1007/s10985-009-9128-2

8. Lindsay B. Composite likelihood methods. *Statistical Inference from Stochastic Processes* 1988, Ed. Prabhu NU. Providence, RI: American Mathematical Society.

9. Varin C, Vidoni P. A note on composite likelihood inference and model selection. *Biometrica* 2005; **92**:519–28. DOI: 10.1093/biomet/92.3.519

10. Varin C. On composite marginal likelihoods. *Advances in Statistical Analysis* 2008; **92**:1–28. DOI: 10.1007/s10182-008-0060-7

11. Robins JM, Hernan MA, Brumback B. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology* 2000; **11**(5):550-560. DOI: 10.1097/00001648-200009000-00011

12. Hernan MA, Brumback B, Robins JM. Marginal Structural Models to Estimate the Causal Effect of Zidovudine on the Survival of HIV-Positive Men. *Epidemiology* 2000; **11**(5):561–70. DOI: 10.1097/00001648-200009000-00012

13. Ledergerber B, Egger M, Opravil M, Telenti A, Hirschel B, Battegay M, Vernazza P, Sudre P, Flepp M, Furrer H, Francioli P, Weber R. Clinical progression and virological failure on highly active antiretroviral therapy in HIV-1 patients: a prospective cohort study. Swiss HIV Cohort Study. *Lancet*. 1999; **353**(9156):863–868. DOI: 10.1016/S0140-6736(99)01122-8

14. Sterne JAC, Hernan MA, Ledergerber B, Tilling K, Weber R, Sendi P, Rickenbach M, Robins JM, Egger M, Swiss HIV Cohort Study. Long-term effectiveness of potent antiretroviral therapy in preventing AIDS and death: a prospective cohort study. *Lancet* 2005; **366**:378–384. DOI: 10.1016/S0140-6736(05)67022-5

15. Røysland K. A martingale approach to continuous time marginal structural models. Arxiv preprint arXiv:0901.2593. To appear in Bernoulli.

16. Andersen PK, Borgan Ø, Gill RD, Keiding N. *Statistical models based on counting processes*. Springer Series in Statistics, Springer-Verlag, 1993.

17. Diggle P, Farewell D, Henderson R. Analysis of longitudinal data with drop-out: objectives, assumptions and a proposal. *Appl. Statist.* 2007; **56**:499–550. DOI: 10.1111/j.1467-9876.2007.00590.x

18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, Chapman & Hall, 1993.

19. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2008. DOI: 10.1007\978-3-540-74686-7

20. Lu W, Andrieu J-M. HIV protease inhibitors restore impaired T-cell proliferative response in vivo and in vitro: a viral-suppression-independent mechanism. *Immunobiology* 2000; **96**(1):250–258.

21. Monini P, Sgadari C, Barillari G, Ensoli B. HIV protease inhibitors: antiretroviral agents with anti-inflammatory, anti-angiogenic and anti-tumour activity. *Journal of Antimicrobial Chemotherapy* 2003; **51**:207–211. DOI: 10.1093/jac/dkg086

22. Sgadari C, Monini P, Barillari G, Ensoli B. *Lancet Oncology* 2003; **4**:537–547. DOI: 10.1016/S1470-2045(03)01192-6

23. Cole SR, Hernan MA. Fallibility in estimating direct effects. *International Journal of Epidemiology* 2002; **31**:163–165. DOI: 10.1093/ije/31.1.163

24. Kaufman JS, MacLehose RF, Kaufman S. A further critique of the analytic strategy of adjusting for covariates to identify biologic mediation. *Epidemiologic Perspectives & Innovations* 2004; **1**(4). DOI: 10.1186/1742-5573-1-4

25. Blakely T. Commentary: Estimating direct and indirect effects - fallible in theory, but in the real world? *International Journal of Epidemiology* 2002; **31**:166–167.